Information Retrieval Models

IR Herbstschule, Dagstuhl, 2008

Thomas Roelleke

Wednesday October 1st, 2008

Introduction & Motivation

- A retrieval model is an application of a mathematical framework to measure
  - the distance between document d and query q
  - the relevance of document d wrt query q
- There are heuristic and — so-called — probabilistic retrieval models
- This seminar is about the theoretical foundations of IR models
- Most models presented here have good and stable performance

Time-line of Retrieval Models: 1960 - 1990

[Maron and Kuhns, 1960]: On Relevance, Probabilistic Indexing, and IR

[Salton, 1971, Salton et al., 1975]: VSM, TF-IDF

[Rocchio, 1971]: Relevance feedback

[Robertson and Sparck Jones, 1976]: BIR

[Croft and Harper, 1979]: BIR without relevance

- [Bookstein, 1980, Salton et al., 1983]: Fuzzy, extended Boolean
- [van Rijsbergen, 1986, van Rijsbergen, 1989]: $P(d \rightarrow q)$
- [Cooper, 1988, Cooper, 1991, Cooper, 1994]: Beyond Boole, ...
- [Dumais et al., 1988, Deerwester et al., 1990]: Latent semantic indexing

Time-line of Retrieval Models: 1990 - ...

[Turtle and Croft, 1990, Turtle and Croft, 1991a]: PIN
[Fuhr, 1992]: Prob Models in IR
[Margulis, 1992]: Poisson
[Robertson and Walker, 1994, Robertson et al., 1995]: 2-Poisson, BM25
[Wong and Yao, 1995]: $P(d \rightarrow q)$
[Brin and Page, 1998, Kleinberg, 1999]: Pagerank and Hits
[Ponte and Croft, 1998, Lavrenko and Croft, 2001]: LM, Relevance-based LM

- [Hiemstra, 2000]: TF-IDF and LM
- [Amati and van Rijsbergen, 2002, He and Ounis, 2005]: DFR
- [Croft and Lafferty, 2003, Lafferty and Zhai, 2003]: LM book
- [Zaragoza et al., 2003]: Bayesian LM
- [Fang and Zhai, 2005]: Axiomatic approach
- [Roelleke and Wang, 2006]: Parallel derivation

Books

[van Rijsbergen, 1979]: online

[Baeza-Yates and Ribeiro-Neto, 1999]

[Grossman and Frieder, 1998, Grossman and Frieder, 2004]: text retrieval and VSM in SQL

[Belew, 2000]: information and noise

[Manning et al., 2008]: Introduction to Information Retrieval

Running Example: Toy collection with 10 documents

term20	
Term	DocId
sailing	doc1
boats	doc1
sailing	doc2
boats	doc2
sailing	doc2
sailing	doc3
east	doc3
coast	doc3
sailing	doc4
boats	doc5
sailing	doc6
boats	doc6
east	doc6
coast	doc6
sailing	doc6
boats	doc6
boats	doc7
coast	doc8
coast	doc9
sailing	doc10

The construction plan of this toy collection is as follows: index "term20" contains 20 entries (tuples) and 10 documents; for relevance feedback (BIR model), 4 out of the 10 documents will be viewed as relevant, and the other 6 will be viewed as non-relevant.

Among the first 10 tuples of term20, there is one reoccurring tuple, namely (sailing,doc2); this tuple is to demonstrate the effect of the within-document term frequency $\mathrm{tf}(t, d)$.

The second half of term20 starts with document "doc6", and and this is a long document to demonstrate the effect of document length normalisation.

Notation

Book's notation	Comment
$n_L(t,d)$ $N_L(d)$	number of *locations* at which term t occurs in document number of *locations* in document d (document length)
$n_D(t,c)$ $N_D(c)$	number of *documents* in which term t occurs in collectio number of *documents* in collection c
$n_L(t,q)$ $N_L(q)$	number of *locations* at which term t occurs in query q number of *locations* in query q (query length)
$n_L(t,c)$ $N_L(c)$	number of *locations* at which term t occurs in collection number of *locations* in collection c ("collection length")
$\mathrm{avgtf\_coll}(t,c) := \frac{n_L(t,c)}{N_D(c)}$ $\mathrm{avgtf\_elite}(t,c) := \frac{n_L(t,c)}{n_D(t,c)}$	average term frequency in documents of collection average term frequency in documents of elite set
$\mathrm{avgdl}(c) := \frac{N_L(c)}{N_D(c)}$ $\mathrm{pivdl}(d,c) := \frac{N_L(d)}{\mathrm{avgdl}(c)}$	average document length ($N_L(d_{\mathrm{avg}})$) pivoted document length

Notation

Probability	Comment
$P_L(t\|d) := \frac{n_L(t,d)}{N_L(d)}$	location-based within-document term probability
$P_L(t\|c) := \frac{n_L(t,c)}{N_L(c)}$	location-based collection-wide term probability
$P_D(t\|c) := \frac{n_D(t,c)}{N_D(c)}$	document-based collection-wide term probability

Notation: Example

$N_L(c)$	20	N
$N_D(c)$	10	
$\mathrm{avgdl}(c)$	20/10=2	

t	sailing	boats	
$n_L(t, c)$	8	6	TF
$n_D(t, c)$	6	5	n_t
$P_L(t\|c)$	8/20	6/20	
$P_D(t\|c)$	6/10	5/10	$\mathrm{df}(t)$
$\mathrm{avgtf\_elite}(t, c)$	8/6	6/5	λ
$\mathrm{avgtf\_coll}(t, c)$	8/10	6/10	λ

TF-IDF Model(s)

1. TF-IDF term weight and TF-IDF RSV
   - TF: within-document term frequency
   - IDF: collection-wide inverse document frequency
2. Example

TF-IDF: TF variants

Definition (TF-IDF term weight)

$$\mathrm{tf}_{\mathrm{total}}(t,d) := n_L(t,d) \tag{1}$$

$$\mathrm{tf}_{\mathrm{sum}}(t,d) := \frac{n_L(t,d)}{N_L(d)} \tag{2}$$

$$\mathrm{tf}_{\mathrm{max}}(t,d) := \frac{n_L(t,d)}{n_L(t_{\mathrm{max}},d)} \tag{3}$$

$$\mathrm{tf}_{\mathrm{piv}}(t,d) := \frac{n_L(t,d)}{n_L(t,d)+K} \tag{4}$$

K? $K_{\mathrm{BM25}} = b \cdot \frac{\mathrm{dl}}{\mathrm{avgdl}} + (1-b)$.

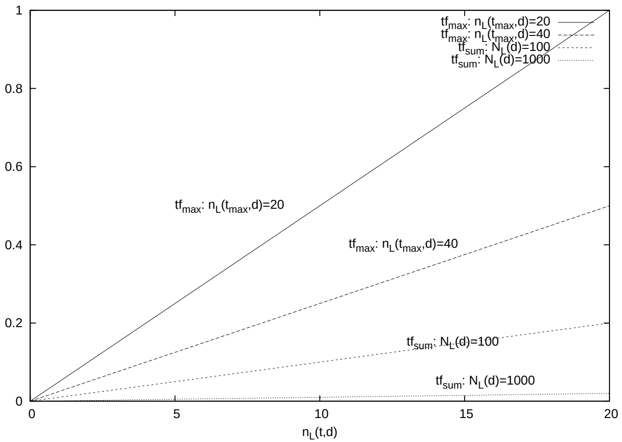
TF-IDF Example: TF variants

tf_sum:10		
$P(t\|d)$	Term	DocId
0.500	sailing	doc1
0.500	boats	doc1
0.667	sailing	doc2
0.333	boats	doc2
0.333	sailing	doc3
0.333	east	doc3
0.333	coast	doc3
1.000	sailing	doc4
1.000	boats	doc5
0.333	sailing	doc6

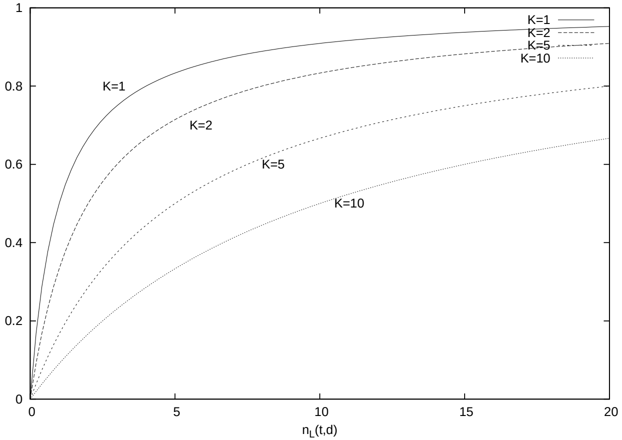
tf_max:10		
$P(t\|d)$	Term	DocId
1.000	sailing	doc1
1.000	boats	doc1
1.000	sailing	doc2
0.500	boats	doc2
1.000	sailing	doc3
1.000	east	doc3
1.000	coast	doc3
1.000	sailing	doc4
1.000	boats	doc5
1.000	sailing	doc6

tf_piv:10		
$P(t\|d)$	Term	DocId
0.500	sailing	doc1
0.500	boats	doc1
0.571	sailing	doc2
0.400	boats	doc2
0.400	sailing	doc3
0.400	east	doc3
0.400	coast	doc3
0.667	sailing	doc4
0.667	boats	doc5
0.400	sailing	doc6

TF-IDF: linear TF curves



TF-IDF: BM25 piv TF curves



TF-IDF: DF and IDF

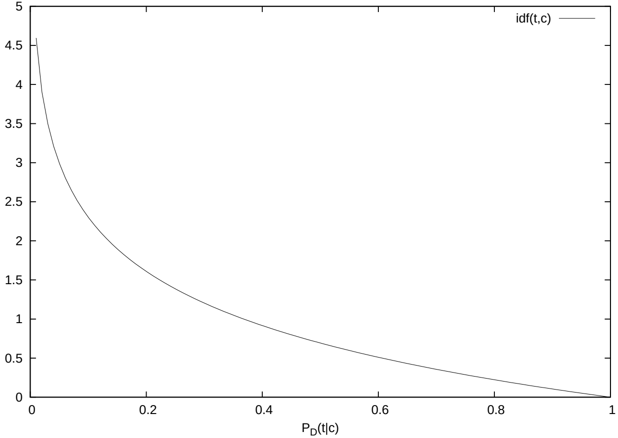
Definition (TF-IDF term weight)

$$\mathrm{df}(t, c) := \frac{n_D(t, c)}{N_D(c)} \tag{5}$$

$$\mathrm{idf}(t, c) := -\log \mathrm{df}(t, c) \tag{6}$$

$$w_{\mathrm{TF-IDF}}(t, d, q, c) := \mathrm{tf}(t, d) \cdot \mathrm{tf}(t, q) \cdot \mathrm{idf}(t, c) \tag{7}$$

TF-IDF: IDF curve



TF-IDF RSV

Definition ($\mathrm{RSV}_{\mathrm{TF-IDF}}$)

$$\mathrm{RSV}_{\mathrm{TF-IDF}}(d, q, c) := \sum_t w_{\mathrm{TF-IDF}}(t, d, q, c) \quad (8)$$

$$= \sum_t \mathrm{tf}(t, d) \cdot \mathrm{tf}(t, q) \cdot \mathrm{idf}(t, c) \quad (9)$$

TF-IDF Example: DF and IDF

df	
$P(t \text{ occurs}\|c)$	Term
0.600	sailing
0.500	boats
0.200	east
0.400	coast

log_df	
$P(t \text{ occurs}\|c)$	Term
0.511	sailing
0.693	boats
1.609	east
0.916	coast

pidf	
$P(t \text{ informs}\|c)$	Term
0.317	sailing
0.431	boats
1.000	east
0.569	coast

$$\mathrm{pidf}(t, c) := P(t \text{ informs}|c) = \mathrm{idf}(t, c)/\mathrm{maxidf}(c) \tag{10}$$

TF-IDF Example: Query term weighting

qterm_idf		
$P(t \text{ informs}\|c)$	Term	QueryId
0.317	sailing	q1
0.431	boats	q1

qterm_norm_idf		
$P(t \text{ informs}\|c)$	Term	QueryId
0.424	sailing	q1
0.576	boats	q1

TF-IDF Example: Retrieval result

tf_sum_idf_retrieve		
RSV	DocId	QueryId
0.431	doc7	q1
0.431	doc5	q1
0.374	doc1	q1
0.355	doc2	q1
0.317	doc10	q1
0.317	doc4	q1
0.249	doc6	q1
0.106	doc3	q1

tf_max_idf_retrieve		
RSV	DocId	QueryId
1.000	doc6	q1
1.000	doc1	q1
0.712	doc2	q1
0.576	doc7	q1
0.576	doc5	q1
0.424	doc10	q1
0.424	doc4	q1
0.424	doc3	q1

tf_piv_idf_retrieve		
RSV	DocId	QueryId
0.500	doc1	q1
0.473	doc2	q1
0.400	doc6	q1
0.384	doc7	q1
0.384	doc5	q1
0.283	doc10	q1
0.283	doc4	q1
0.170	doc3	q1

TF-IDF Example: RSV computation

$$\begin{aligned}
\mathrm{RSV}_{\mathrm{TF_sum-IDF}}(\mathrm{doc7}) &= 0.431 = 1.0 \cdot 0.431 \\
\mathrm{RSV}_{\mathrm{TF_sum-IDF}}(\mathrm{doc1}) &= 0.374 = 0.5 \cdot 0.317 + 0.5 \cdot 0.431
\end{aligned}$$

$$\begin{aligned}
\mathrm{RSV}_{\mathrm{TF_piv-IDF}}(\mathrm{doc1}) &= 0.5 = \frac{1}{1+2/2} \cdot 0.424 + \frac{1}{1+2/2} \cdot 0.576 \\
\mathrm{RSV}_{\mathrm{TF_piv-IDF}}(\mathrm{doc6}) &= 0.4 = \frac{2}{2+6/2} \cdot 0.424 + \frac{2}{2+6/2} \cdot 0.576 \\
\mathrm{RSV}_{\mathrm{TF_piv-IDF}}(\mathrm{doc7}) &= 0.384 = \frac{1}{1+1/2} \cdot 0.576
\end{aligned}$$

BIR Model

1. Background
2. BIR term weight and BIR RSV
3. Example

BIR Background

[Robertson and Sparck Jones, 1976]

Derivation: Start from probabilistic odds:

$$O(r|d,q) := \frac{P(r|d,q)}{P(\bar{r}|d,q)} \tag{11}$$

The application of Bayes theorem, a term independence assumption, and a non-query term assumption lead to the BIR term weight and BIR RSV.

BIR term weight

Definition (BIR term weight)

The BIR term weight is:

$$w_{\mathrm{BIR}}(t,q) := \frac{P(t|r)}{P(t|\bar{r})} \cdot \frac{P(\bar{t}|\bar{r})}{P(\bar{t}|r)} \tag{12}$$

The simplified form considers term presence only:

$$w_{\mathrm{BIR-F1}}(t,q) := \frac{P(t|r)}{P(t|\bar{r})} \tag{13}$$

BIR RSV

Definition (RSV_{BIR})

$$\text{RSV}_{\text{BIR}}(d,q) := \sum_{t \in d \cap q} \log w_{\text{BIR}}(t,q) \tag{14}$$

BIR: Term presence and absence

Definition (Variants of the BIR term weight)

	$\bar{r} = c$	$\bar{r} = c \setminus r$
Presence only	$\frac{r_t/R}{n_t/N}$	$\frac{r_t/R}{(n_t - r_t)/(N-R)}$
Presence and absence	$\frac{r_t/(R-r_t)}{n_t/(N-n_t)}$	$\frac{r_t/(R-r_t)}{(n_t-r_t)/(N-R-(n_t-r_t))}$

BIR: Zero probability term

Definition (Variants of the BIR term weight)

	$\bar{r} = c$	$\bar{r} = c \setminus r$
Presence only	$\frac{(r_t+0.5)/(R+1)}{(n_t+1)/(N+2)}$	$\frac{(r_t+0.5)/(R+1)}{(n_t-r_t+0.5)/(N-R+1)}$
Presence and absence	$\frac{(r_t+0.5)/(R-r_t+0.5)}{(n_t+1)/(N-n_t+1)}$	$\frac{(r_t+0.5)/(R-r_t+0.5)}{(n_t-r_t+0.5)/(N-R-(n_t-r_t)+0.5)}$

BIR Example

qterm

Term	DocId
sailing	q1
boats	q1

relevant

QueryId	DocId
q1	doc2
q1	doc4
q1	doc6
q1	doc8

non_relevant

QueryId	DocId
q1	doc1
q1	doc3
q1	doc5
q1	doc7
q1	doc9
q1	doc10

BIR Example: index of relevant and non-relevant documents

relColl		
Term	DocId	QueryId
sailing	doc2	q1
boats	doc2	q1
sailing	doc2	q1
sailing	doc4	q1
sailing	doc6	q1
boats	doc6	q1
east	doc6	q1
coast	doc6	q1
sailing	doc6	q1
boats	doc6	q1
coast	doc8	q1

non_relColl		
Term	DocId	QueryId
sailing	doc1	q1
boats	doc1	q1
sailing	doc3	q1
east	doc3	q1
coast	doc3	q1
boats	doc5	q1
boats	doc7	q1
coast	doc9	q1
sailing	doc10	q1

BIR Example: The trick with the virtual doc

relColl_virtual		
Term	DocId	QueryId
sailing	doc2	q1
boats	doc2	q1
sailing	doc2	q1
sailing	doc4	q1
sailing	doc6	q1
boats	doc6	q1
east	doc6	q1
coast	doc6	q1
sailing	doc6	q1
boats	doc6	q1
coast	doc8	q1
sailing	virtualDoc	q1
boats	virtualDoc	q1

non_relColl_virtual		
Term	DocId	QueryId
sailing	doc1	q1
boats	doc1	q1
sailing	doc3	q1
east	doc3	q1
coast	doc3	q1
boats	doc5	q1
boats	doc7	q1
coast	doc9	q1
sailing	doc10	q1
sailing	virtualDoc	q1
boats	virtualDoc	q1

The trick: add the query to the set of relevant and non-relevant documents

Guarantees $P(t|r) > 0$ and $P(t|\bar{r}) > 0$

BIR Example: Term probabilities

term_r

$P(t \mid r)$	Term	QueryId
0.800	sailing	q1
0.600	boats	q1
0.200	east	q1
0.400	coast	q1

term_not_r

$P(t \mid \bar{r})$	Term	QueryId
0.571	sailing	q1
0.571	boats	q1
0.143	east	q1
0.286	coast	q1

term_c

$P(t \mid c)$	Term
0.600	sailing
0.500	boats
0.200	east
0.400	coast

bir_term_weight

	Term	QueryId
1.400	sailing	q1
1.050	boats	q1
1.400	east	q1
1.400	coast	q1

bir_c_term_weight

	Term	QueryId
1.333	sailing	q1
1.200	boats	q1
1.000	east	q1
1.000	coast	q1

BIR Example: Term weight computation

$$
\begin{aligned}
w_{\mathrm{BIR}}(\text{sailing}, q) &= 1.40 = \frac{0.8}{0.571} \\
w_{\mathrm{BIR}}(\text{boats}, q) &= 1.05 = \frac{0.6}{0.571} \\
w_{\mathrm{BIR}\,c}(\text{sailing}, q) &= 1.333 = \frac{0.8}{0.6} \\
w_{\mathrm{BIR}\,c}(\text{boats}, q) &= 1.20 = \frac{0.6}{0.5}
\end{aligned}
$$

BIR Example: Retrieval results

bir_retrieve		
RSV_{BIR}	DocId	QueryId
1.470	doc6	q1
1.470	doc2	q1
1.470	doc1	q1
1.400	doc10	q1
1.400	doc4	q1
1.400	doc3	q1
1.050	doc7	q1
1.050	doc5	q1

bir_c_retrieve		
RSV_{BIR}	DocId	QueryId
1.600	doc6	q1
1.600	doc2	q1
1.600	doc1	q1
1.333	doc10	q1
1.333	doc4	q1
1.333	doc3	q1
1.200	doc7	q1
1.200	doc5	q1

BIR Example: RSV computation

$$\begin{aligned} \mathrm{RSV}_{\mathrm{BIR}}(\mathrm{doc1}) &= 1.470 = 1.40 \cdot 1.05 \\ \mathrm{RSV}_{\mathrm{BIR\_c}}(\mathrm{doc1}) &= 1.60 = 1.333 \cdot 1.20 \end{aligned}$$

Poisson Model

1. Background
2. Binomial probability
3. Poisson probability (approximation of Binomial prob)
4. Analogy between $P(n \text{ sunny days})$ and $P(n_L(t, d) \text{ locations})$
5. Poisson term weight and Poison RSV
6. Example

Poisson Background

[Margulis, 1992]: N-dimensional Poisson

[Church and Gale, 1995]: idf is deviation from Poisson

[Robertson and Walker, 1994]: 2-Poisson model

Binomial probability

Definition (Binomial probability)

$$P_{\text{Binomial}}(k_t|c) := \binom{N}{k_t} \cdot p_t^{k_t} \cdot (1 - p_t)^{(N-k_t)} \tag{15}$$

For example, the probability that $k_t = 4$ sunny days occur in $N = 7$ days; the single event probability is $p_t = \frac{180}{360} = 0.5$.

$$P_{\text{Binomial}}(k_t = 4|c) = \binom{7}{4} \cdot 0.5^4 \cdot (1 - 0.5)^{7-4} \approx 0.2734 \tag{16}$$

Poisson probability

Definition (Poisson probability)

$$P_{\text{Poisson}}(k_t|c) := \frac{(\lambda(t,c))^{k_t}}{k_t!} \cdot e^{-\lambda(t,c)} \tag{17}$$

For example, the probability that $k_t = 4$ sunny days occur in a week; the average is $180/360 * 7 = 3.5$ sunny days per week.

$$P_{\text{Poisson}}(k_t = 4|c) = \frac{(3.5)^4}{4!} \cdot e^{-3.5} \approx 0.1888 \tag{18}$$

Analogy of Days/Holiday and Locations/Document

Event space	Days	Locations
k_t	sunny days	term locations
trial sequence	holiday h sequence of days	document d sequence of locations
background model	year y	collection c
N: number of trials, i.e. length of sequence	days in holiday: $N_{\text{Days}}(h)$	locations in document: $N_{\text{Locations}}(d)$
single event probability	$P_{\text{Days}}(\text{sunny}\|y) := \frac{n_{\text{Days}}(\text{sunny},y)}{N_{\text{Days}}(y)}$	$P_{\text{Locations}}(t\|c) := \frac{n_{\text{Locations}}(t,c)}{N_{\text{Locations}}(c)}$

Poisson term weight

Definition (Poisson term weight)

The Poisson term weight is:

$$w_{\mathrm{Poisson}}(t, d, r, \bar{r}) := \left(\frac{\lambda(t, r)}{\lambda(t, \bar{r})}\right)^{n_L(t,d)} \tag{19}$$

Poisson RSV

Definition ($\mathrm{RSV}_{\mathrm{Poisson}}$)

$$\mathrm{RSV}_{\mathrm{Poisson}}(d, q, r, \bar{r}) := \sum_{t \in d \cap q} \log w_{\mathrm{Poisson}}(t, d, r, \bar{r}) \tag{20}$$

$$= \sum_{t \in d \cap q} n_L(t, d) \cdot -\log \lambda(t, \bar{r}) - n_L(t, d) \cdot -\log \lambda(t, r) \tag{21}$$

$$= \sum_{t \in d \cap q} n_L(t, d) \cdot -\log \frac{n_L(t, \bar{r})}{N_D(\bar{r})} - \tag{22}$$

$$\sum_{t \in d \cap q} n_L(t, d) \cdot -\log \frac{n_L(t, r)}{N_D(r)}$$

2-Poisson Model

[Robertson and Walker, 1994]

...

BM25 Model

[Robertson et al., 1995]: Okapi/BM25

BM25 tutorials SIGIR 2007 and 2008: Hugo Zaragoza, Stephen Robertson

BM25 term weight

Definition (BM25 term weight)

$$w_{\mathrm{BM25}}(t, d, q) := \frac{\mathrm{tf}'}{\mathrm{tf}' + k_1} \cdot w_{\mathrm{BIR}}(t, q) \cdot \frac{\mathrm{qtf}}{\mathrm{qtf} + k_3} \tag{23}$$

$$\mathrm{tf}' := \frac{\mathrm{tf}}{b \cdot \frac{\mathrm{dl}}{\mathrm{avgdl}} + (1 - b)} \tag{24}$$

BM25 term RSV

Definition ($\mathrm{RSV}_{\mathrm{BM25}}$)

$$\mathrm{RSV}_{\mathrm{BM25}}(d, q) := \left[\sum_{t \in d \cap q} w_{\mathrm{BM25}}(t, d, q)\right] + k_2 \cdot \mathrm{ql} \cdot \frac{\mathrm{avgdl} - \mathrm{dl}}{\mathrm{avgdl} + \mathrm{dl}} \tag{25}$$

BM25 notation

traditional notation	book notation	comment
tf	$n_L(t,d)$	within-document term frequency
tf'	$\frac{n_L(t,d)}{b \cdot \frac{N_L(d)}{\mathrm{avgdl}(c)} + (1-b)}$	normalised within-document term frequency (pivoted document length $\mathrm{pivdl}(d,c) := \frac{N_L(d)}{\mathrm{avgdl}(c)}$)
qtf	$n_L(t,q)$	within-query term frequency
b	b	constant to adjust impact of document length normalisation
k_1	k_1	constant to adjust impact of tf
ql	$N_L(q)$	query length: locations in query q
dl	$N_L(d)$	document length: locations in document d
avgdl	$\mathrm{avgdl}(c)$	average document length; also $N_L(d_{\mathrm{avg}})$
$w_t^{(1)}$	$w_{\mathrm{BIR}}(t,q)$	BIR term weight
k_2	k_2	constant to adjust impact of document length
k_3	k_3	constant to adjust impact of qtf

Language Modelling (LM)

1. Background
2. LM term weight and LM RSV
3. Example

LM Background

[Ponte and Croft, 1998, Lavrenko and Croft, 2001]: LM, Relevance-based LM

[Hiemstra, 2000]: A probabilistic justification for using tf.idf term weighting in information retrieval

[Croft and Lafferty, 2003]: Language Modeling for Information Retrieval

Victor Lavrenko LM tutorial SIGIR 2003

[Zaragoza et al., 2003]: A Bayesian ...

LM term weight

Definition (LM term weight)

For the within-document term probability $P(t|d)$ and the collection-wide term probability $P(t|c)$, the linear mixture is:

$$P(t|d,c) := \delta \cdot P(t|d) + (1 - \delta) \cdot P(t|c) \tag{26}$$

LM RSV

Definition ($\mathrm{RSV}_{\mathrm{LM}}$)

$$\mathrm{RSV}_{\mathrm{LM}}(d, q, c) := \log P(q|d, c) = \sum_{t \in q} \log P(t|d, c) \tag{27}$$

LM Example: document and collection/background model

docModel		
$P(t\|d)$	Term	DocId
0.500000	sailing	doc1
0.500000	boats	doc1
0.666667	sailing	doc2
0.333333	boats	doc2
0.333333	sailing	doc3
0.333333	east	doc3
0.333333	coast	doc3
1.000000	sailing	doc4
1.000000	boats	doc5
0.333333	sailing	doc6
0.333333	boats	doc6
0.166667	east	doc6
0.166667	coast	doc6
1.000000	boats	doc7
1.000000	east	doc8
1.000000	coast	doc9
1.000000	sailing	doc10

collModel	
$P(t\|c)$	Term
0.400000	sailing
0.300000	boats
0.150000	east
0.150000	coast

LM Example: Term weights/probabilities

lm1_term_weight:20		
$P(t\|d,c)$	Term	DocId
0.480000	sailing	doc1
0.460000	boats	doc1
0.613333	sailing	doc2
0.326667	boats	doc2
0.346667	sailing	doc3
0.286667	east	doc3
0.306667	coast	doc3
0.880000	sailing	doc4
0.860000	boats	doc5
0.346667	sailing	doc6
0.326667	boats	doc6
0.153333	east	doc6
0.173333	coast	doc6
0.860000	boats	doc7
0.800000	coast	doc8
0.800000	coast	doc9
0.880000	sailing	doc10
0.080000	sailing	doc5
0.080000	sailing	doc7
0.060000	boats	doc3

The following table illustrates for some term-document tuples in relation "lm1_term_weight" the computation of the mixed probabilities (mixture parameter $\delta = 0.8$).

lm1_term_weight		
$P(t\|d,c)$	Term	DocId
$0.48 = 0.8 \cdot 0.5 + 0.2 \cdot 0.4$	sailing	doc1
$0.46 = 0.8 \cdot 0.5 + 0.2 \cdot 0.3$	boats	doc1
$0.61333 = 0.8 \cdot 0.667 + 0.2 \cdot 0.4$	sailing	doc2
$0.32667 = 0.8 \cdot 0.333 + 0.2 \cdot 0.3$	boats	doc2
...	...	...

LM Example: Retrieval results

lm1_retrieve		
$P(q\|d,c)$	DocId	QueryId
0.220800	doc1	q1
0.200356	doc2	q1
0.113244	doc6	q1
0.068800	doc7	q1
0.068800	doc5	q1
0.052800	doc10	q1
0.052800	doc4	q1
0.020800	doc3	q1

For example, the computation of the probabilities of “doc1” and “doc2” is as follows:

$$P(q|\text{doc1}, c) =$$
$$= P(\text{sailing}|\text{doc1}, c) \cdot P(\text{boats}|\text{doc1}, c)$$
$$= 0.48 \cdot 0.46 = 0.2208$$

$$P(q|\text{doc2}, c) =$$
$$= P(\text{sailing}|\text{doc2}, c) \cdot P(\text{boats}|\text{doc2}, c)$$
$$= 0.6133 \cdot 0.3266 = 0.2003$$

More Models

1. Probabilistic Inference Network (PIN) Model
2. Divergence from Randomness (DFR) Model
3. Link-based Models (TF boosting, page-rank)
4. Classification-oriented Models (Bayesian, Support-vector machine (SVM))
5. Relevance feedback models (Rocchio, ...)
6. More “models”

Probabilistic Inference Network (PIN) Model

1. Background
2. PIN term weight and PIN RSV
3. Example

Background

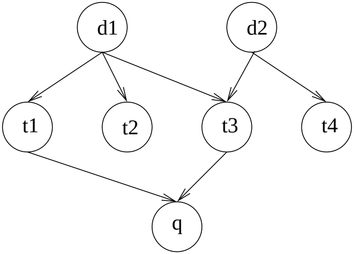
[Turtle and Croft, 1990, Turtle and Croft, 1991a, Turtle and Croft, 1991b]: PIN for Document Retrieval, Efficient Prob Inference for Text Retrieval, Evaluation of an PIN-based Retrieval Model (evolution: document, text, model)

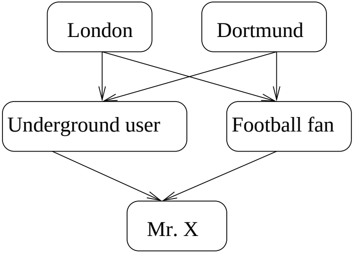
[Croft and Turtle, 1992]: Retrieval of complex objects (EDBT)

[Turtle and Croft, 1992]: A comparison of text retrieval models (CJ)

[Metzler and Croft, 2004]: Combining LM and PIN (IP&M)

PIN's: Document retrieval and "Find Mr. X"





Link Matrix

$$P(q|d) = \sum_{x} P(q|x) \cdot P(x|d) \tag{28}$$

$$\begin{pmatrix} P(q|d) \\ P(\bar{q}|d) \end{pmatrix} = L \cdot \begin{pmatrix} P(x_1|d) \\ \vdots \\ P(x_n|d) \end{pmatrix} \tag{29}$$

Link Matrices L_{or} and L_{and}

$$L_{\text{or}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{30}$$

$$L_{\text{and}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{31}$$

Link Matrix for Closed Form with $O(n)$

$$L = \begin{bmatrix} 1 & \frac{w_1+w_2}{w_0} & \frac{w_1+w_3}{w_0} & \frac{w_1}{w_0} & \frac{w_2+w_3}{w_0} & \frac{w_2}{w_0} & \frac{w_3}{w_0} & 0 \\ 0 & \frac{w_3}{w_0} & \frac{w_2}{w_0} & \frac{w_2+w_3}{w_0} & \frac{w_1}{w_0} & \frac{w_1+w_3}{w_0} & \frac{w_1+w_2}{w_0} & 1 \end{bmatrix} \quad (32)$$

$w_0 = \sum_i w_i$

$$\frac{w_1}{w_0} \cdot P(t_1|d) + \frac{w_2}{w_0} \cdot P(t_2|d) + \frac{w_3}{w_0} \cdot P(t_3|d) \quad (33)$$

PIN term weight

Definition (PIN term weight)

$$w_{\text{PIN}}(t, d, q) := \frac{P(q|t) \cdot P(t|d)}{\sum_t P(q|t)} \tag{34}$$

Probabilistic (PIN) interpretation of TF-IDF?

PIN RSV

Definition ($\mathrm{RSV}_{\mathrm{PIN}}$)

$$\mathrm{RSV}_{\mathrm{PIN}}(d, q) := \sum_t w_{\mathrm{PIN}}(t, d, q) \tag{35}$$

$$= \frac{1}{\sum_t P(q|t)} \cdot \sum_t P(q|t) \cdot P(t|d) \tag{36}$$

DFR: Divergence from Randomness

"The more the divergence of the within-document term frequency from its frequency within the collection, the more divergent from randomness the term is, meaning the more the information carried by the term in the document."

[Amati and Rijsbergen, 2002, Amati and van Rijsbergen, 2002]: Pareto (ECIR), measuring the DFR (TOIS)

Link-based Models

1. TF-boosting
2. Page-rank

TF-boosting

$$n_{L,\text{boosted}}(t, d) := n_L(t, d) + \sum_a \text{link(a,d)} \cdot n_L(t, a) \qquad (37)$$

[Craswell et al., 2001]: Effective site finding using link anchor information

Page-rank

$$\text{page-rank}(y) := d + (1 - d) \cdot \sum_{x} \text{link}(x, y) \cdot \frac{\text{page-rank(x)}}{N(x)} \quad (38)$$

[Brin and Page, 1998, Kleinberg, 1999]

Classification-oriented Models

1. Bayesian classifier
2. Support-vector machine (SVM)

[Joachims, 2000, Klinkenberg and Joachims, 2000]: Generalisation performance, Concept Drift with SVM

[Sebastiani, 2002]: Machine-learning in automated text categorisation

Trend: Learning to rank

Classifcation: Bayesian Classifier

feature independence assumption: $P(c|d) = \prod_{t \in c} P(t|d)$ (39)

Classifcation: Support-vector Machine (SVM)

$$\vec{y} = A \cdot \vec{x} + \vec{b} \tag{40}$$

The matrix A is estimated/learned from a set of input-output pairs $(\vec{x}_i, \vec{y}_i)$. The estimation is based on the minimum of the error $\mathrm{err}(A)$. The error can be based on the sum of the squares of $A \cdot \vec{x}_i - \vec{y}_i$ (method of least square polynomials, described in any math text book).

$$\mathrm{err}(A) = \sum_i (A \cdot \vec{x}_i - \vec{y}_i)^2 \tag{41}$$

More “models”

- Boolean model
- Extended Boolean model
- Fuzzy model
- Vector-space “model” (VSM)
- Logical retrieval “model”: $P(d \rightarrow q)$
- Relevance feedback models
- Latent semantic indexing

Relevance Feedback

A classic: [Rocchio, 1966, Rocchio, 1971]:

$$\vec{q}_{\text{revised}} = \alpha \cdot \vec{q}_{\text{initial}} + \beta \cdot \frac{1}{|R|} \sum_{d \in R} \vec{d} - \gamma \cdot \frac{1}{|NR|} \sum_{d \in NR} \vec{d} \qquad (42)$$

The revised query is derived from the initial query, the centroid of relevant documents (set R), and the centroid of non-relevant documents (set NR). The parameters α, β, γ adjust the impact and normalisation of each component.

Relevance Feedback

BIR and BM25 (probabilistic odds) consider relevance feedback data. TF-IDF and LM do not.

Relationships between Retrieval Models

- Vector-space Model (VSM) and Generalised VSM (GVSM)
- $P(d \rightarrow q)$: The probability that d implies q
- $P(r|d,q)$: The probability of relevance
- Parallel derivation

Vector-space Model (VSM): Background

1. The milestone "model" in the 60/70s (SMART system)
2. Replaced Boolean retrieval; stable and good quality of ranking results
3. Approach: Apply vector algebra (cosine) to measure the distance between document and query
4. Estimation of vector components: TF-IDF

VSM: Cosine-based $\mathrm{RSV}_{\mathrm{VSM}}$

$$\cos(\angle(\vec{d},\vec{q})) := \frac{\vec{d}\cdot\vec{q}}{\sqrt{\vec{d}^2}\cdot\sqrt{\vec{q}^2}} \tag{43}$$

$$\mathrm{RSV}_{\mathrm{VSM}}(d,q) := \cos(\angle(\vec{d},\vec{q}))\cdot\sqrt{\vec{q}^2} = \frac{\vec{d}\cdot\vec{q}}{\sqrt{\vec{d}^2}} \tag{44}$$

Generalised Vector-space Model (GVSM)

1. VSM only associates same dimensions/terms
2. GVSM associates different dimensions/terms
   - solve syntactic mismatch problem of semantically related terms
   - query for “classification” ... retrieve documents that contain “categorisation”

GVSM RSV

$$\mathrm{RSV}_{\mathrm{GVSM}}(d, q, G) := \vec{d}^T \cdot G \cdot \vec{q} \tag{45}$$

Identity matrix $G = I$ and scalar product $\vec{d} \cdot \vec{q}$:

$$\vec{d}^T \cdot I \cdot \vec{q} = \vec{d} \cdot \vec{q} = w_{d,1} \cdot w_{q,1} + \ldots + w_{d,n} \cdot w_{q,n} \tag{46}$$

GVSM: Example

$$G = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathrm{RSV}_{\mathrm{GSVM}}(d, q, G) = (w_{d,1} + w_{d,2}) \cdot w_{q,1} + \ldots + w_{d,n} \cdot w_{q,n} \quad (47)$$

The GVSM is useful for matching semantically related terms. For example, let $t_1 =$ "*classification*" and $t_2 =$ "*categorisation*" be two dimensions of the vector space. Then, for the example matrix G above, a query for "classification" ($w_{q,1} = 1$) retrieves a document containing "categorisation" ($w_{d,2} = 1$), even though $w_{q,2} = 0$, i.e. "categorisation" does not occur in the query, and $w_{d,1} = 0$, i.e. "classification" does not occur in the document.

General Matrix Framework: Content-based Retrieval

DT_c: Document-Term matrix of collection c

$TD_c = \text{transpose}(DT_c)$

Term \ Doc	doc1	doc2	doc3	doc4	doc5	$n_D(t, c)$	$n_L(t, c)$
sailing	1	2	1	1		4	5
boats	1	1			1	3	3
east			1			1	1
coast			1			1	1
$n_T(d, c)$	2	2	3	1	1		
$n_L(d, c)$	2	3	3	1	1		

General Matrix Framework: Content-based Retrieval

Content-based document retrieval:

$$\mathrm{RSV}(\vec{d}, \vec{q}) = DT_c \cdot \vec{q} \tag{48}$$

$$\begin{aligned} \text{document similarity: } DD_c &= DT_c \cdot TD_c && (49) \\ \text{term co-occurrence: } TT_c &= TD_c \cdot DT_c && (50) \end{aligned}$$

$$\mathrm{RSV}(\vec{d}, \vec{q}) = DT_c \cdot G \cdot \vec{q} \tag{51}$$

General Matrix Framework: Structure-based Retrieval

PC_c: Parent-Child matrix of collection c

$CP_c = \text{transpose}(PC_c)$

Child \ Parent	doc1	doc2	doc3	doc4	$n_C(d, c)$	$n_L(t, c)$
doc1		1	2		2	3
doc2				1	1	1
doc3					0	0
doc4					0	0
$n_P(d, c)$	0	1	1	1		
$n_L(d, c)$	0	1	2	1		

General Matrix Framework: Structure-based Retrieval

$$\text{parent similarity (co-reference): } PP_c = PC_c \cdot CP_c \quad (52)$$

$$\text{child similarity (co-citation): } CC_c = CP_c \cdot PC_c \quad (53)$$

Exploitation of analogies/dualities between

1. content-based and structure-based retrieval
2. collection space (DT_c, PC_c) and document space (ST_d).

[Roelleke et al., 2006]

$P(d \longrightarrow q)$

- View $P(d \rightarrow q)$ as a measure of relevance [van Rijsbergen, 1986, van Rijsbergen, 1989, Nie, 1992, Meghini et al., 1993, Crestani and van Rijsbergen, 1995]: logical approach good for "semantic" retrieval
- Different interpretations of $P(d \rightarrow q)$ explain traditional IR models (VSM, coordination-level match) [Wong and Yao, 1995]: For $P(q|d)$ set $P(q|t)$ and $P(t|d)$

$$P(q|d) = \sum_t P(t|d) \cdot P(q|t) = \vec{d} \cdot \vec{q}$$

$P(r|d,q)$: The Probability of Relevance

The Bayesian equation $P(h|e) = \frac{P(h,e)}{P(e)}$ is the starting point to estimate the probability $P(r|d,q)$ of relevance, given a document-query pair (d,q).

$$P(r|d,q) = \frac{P(d,q,r)}{P(d,q)} \tag{54}$$

Decomposition of $P(d, q, r)$

The conjunctive probability $P(d, q, r)$ can be decomposed into two products:

$$P(d,q,r) = P(d|q,r) \cdot P(q|r) \cdot P(r) \quad (55)$$

$$= P(q|d,r) \cdot P(d|r) \cdot P(r) \quad (56)$$

In the first product, d depends on q, whereas in the second product, q depends on d.

Term Independence Assumption

The next step views the events d and q as conjunctions of terms. The term events are assumed to be *independent*. Then, the probabilities $P(d|q, r)$ and $P(q|d, r)$ can be decomposed as follows:

$$P(d|q, r) = \prod_{t \in d} P(t|q, r) \quad (57)$$

$$P(q|d, r) = \prod_{t \in q} P(t|d, r) \quad (58)$$

Probabilistic Odds

probabilistic odds: $$O(r|d,q) = \frac{P(r|d,q)}{P(\bar{r}|d,q)} \qquad (59)$$

For documents that are more likely to be relevant than not relevant, $P(r|d,q) > P(\bar{r}|d,q)$, i.e. $O(r|d,q) > 1$.

Estimation of Term Probabilities

Document-based (BIR model):

$$P_D(t|c) = \frac{n_D(t,c)}{N_D(c)} \tag{60}$$

Location-based (LM):

$$P_L(t|c) = \frac{n_L(t,c)}{N_L(c)} \tag{61}$$

Frequency-based (Poisson):

$$P(t|c) = P_{\text{Poisson}}(k_t|c) = \frac{\lambda(t,c)^{k_t}}{k_t!} \cdot e^{-\lambda(t,c)} \tag{62}$$

Parallel Derivation of IR Models

retrieval model	BIR	Poisson	LM
	Presence of terms in $N_D(c)$ Documents	Frequency of terms Locations/Documents	Terms at $N_L(c)$ Locations
term statistics	$n_D(t, c)$	$\lambda = n_L(t, c)/n_D(t, c)$	$n_L(t, c)$
event space	$x_t \in \{0, 1\}$	$k_t \in \{0, 1, \dots, n\}$	$t \in \{t_1, \dots, t_n\}$
term probability	$P(x_t\|c) = n_D(t, c)/N_D(c)$ probability that term t occurs in a document of set c	$P(k_t\|c) = P_{\text{Poisson}, \lambda}(k_t)$ probability that term t occurs k_t times given average occurence λ	$P(t\|c) = n_L(t, c)/N_L(c)$ probability that term t occurs in set c of locations

[Robertson, 2004]: IDF: On theoretical arguments

[Robertson, 2005]: Event spaces

[Roelleke and Wang, 2006]: Parallel derivation

Summary

1. TF-IDF, BIR, Poisson, BM25, LM
2. More models:
   1. PIN, DFR
   2. Link-based Models: TF-boosting, Page-rank
   3. Classification-oriented Models: Bayesian, SVM
   4. More models
3. Relationships between Retrieval Models
   1. VSM and GVSM
   2. $P(d \rightarrow q)$: Probability of d implies q
   3. $P(r|d, q)$: Probability of relevance
   4. Parallel derivation of IR models

Amati, G. and Rijsbergen, C. J. (2002).
Term frequency normalization via Pareto distributions.
In Crestani, F., Girolami, M., and Rijsbergen, C. J., editors, *24th BCS-IRSG European Colloquium on IR Research, Glasgow, Scotland*.

Amati, G. and van Rijsbergen, C. J. (2002).
Probabilistic models of information retrieval based on measuring the divergence from randomness.
ACM Transaction on Information Systems (TOIS), 20(4):357–389.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999).
Modern Information Retrieval.
Addison Wesley.

Belew, R. K. (2000).
Finding out about.
Cambridge University Press.

Bookstein, A. (1980).
Fuzzy requests: An approach to weighted Boolean searches.
Journal of the American Society for Information Science, 31:240–247.

Brin, S. and Page, L. (1998).
The anatomy of a large-scale hypertextual web search engine.
Computer Networks, 30(1-7):107–117.

Church, K. and Gale, W. (1995).
Inverse document frequency (idf): A measure of deviation from poisson.
In *Proceedings of the Third Workshop on Very Large Corpora*, pages 121–130.

Cooper, W. (1991).
Some inconsistencies and misnomers in probabilistic IR.
In Bookstein, A., Chiaramella, Y., Salton, G., and Raghavan, V., editors, *Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 57–61, New York.

Cooper, W. S. (1988).
Getting beyond Boole.
Information Processing and Management, 24(3):243–248.

Cooper, W. S. (1994).
Triennial acm sigir award presentation and paper: The formalism of probability theory in ir: A foundation for an encumbrance.
In [Croft and van Rijsbergen, 1994], pages 242–248.

Craswell, N., Hawking, D., and Robertson, S. E. (2001).
Effective site finding using link anchor information.
In *SIGIR*, pages 250–257.

Crestani, F. and van Rijsbergen, C. J. (1995).
Probability kinematics in information retrieval.
In Fox, E., Ingwersen, P., and Fidel, R., editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 291–299, New York. ACM.

Croft, B. and Lafferty, J., editors (2003).
Language Modeling for Information Retrieval.
Kluwer.

Croft, W. and Harper, D. (1979).

Using probabilistic models of document retrieval without relevance information.
Journal of Documentation, 35:285–295.

Croft, W. and Turtle, H. (1992).
Retrieval of complex objects.
In Pirotte, A., Delobel, C., and Gottlob, G., editors, *Advances in Database Technology — EDBT'92*, pages 217–229, Berlin et al. Springer.

Croft, W. B. and van Rijsbergen, C. J., editors (1994).
Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, London, et al. Springer-Verlag.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990).
Indexing by latent semantic analysis.
Journal of the American Society for Information Science, 41(6):391–407.

Dumais, S. T., Furnas, G. W.and Landauer, T. K., and Deerwester, S. (1988).
Using latent semantic analysis to improve information retrieval.
pages 281–285.

Fang, H. and Zhai, C. (2005).
An exploration of axiomatic approaches to information retrieval.
In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 480–487, New York, NY, USA. ACM.

Fuhr, N. (1992).
Probabilistic models in information retrieval.
The Computer Journal, 35(3):243–255.

Grossman, D. A. and Frieder, O. (1998).
Information Retrieval: Algorithms and Heuristics.
Kluwer, Massachusetts.

Grossman, D. A. and Frieder, O. (2004).
Information Retrieval. Algorithms and Heuristics, 2nd ed., volume 15 of *The Information Retrieval Series*.
Springer.

He, B. and Ounis, I. (2005).
Term frequency normalisation tuning for BM25 and DFR models.
In *ECIR*, pages 200–214.

Hiemstra, D. (2000).
A probabilistic justification for using tf.idf term weighting in information retrieval.
International Journal on Digital Libraries, 3(2):131–139.

Joachims, T. (2000).
Estimating the generalization performance of an svm efficiently.
In [Langley, 2000], pages 431–438.

Kleinberg, J. (1999).
Authoritative sources in a hyperlinked environment.
Journal of ACM, 46.

Klinkenberg, R. and Joachims, T. (2000).
Detecting concept drift with support vector machines.
In [Langley, 2000], pages 487–494.

Lafferty, J. and Zhai, C. (2003).

Probabilistic Relevance Models Based on Document and Query Generation, chapter 1.
In [Croft and Lafferty, 2003].

Langley, P., editor (2000).
Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Standord, CA, USA, June 29 - July 2, 2000. Morgan Kaufmann.

Lavrenko, V. and Croft, W. B. (2001).
Relevance-based language models.
In *SIGIR*, pages 120–127.

Manning, C. D., Raghavan, P., and Schuetze, H., editors (2008).
Introduction to Information Retrieval.
Cambridge University Press.

Margulis, E. (1992).
N-poisson document modelling.
In Belkin, N., Ingwersen, P., and Pejtersen, M., editors, *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 177–189, New York.

Maron, M. and Kuhns, J. (1960).
On relevance, probabilistic indexing, and information retrieval.
Journal of the ACM, 7:216–244.

Meghini, C., Sebastiani, F., Straccia, U., and Thanos, C. (1993).
A model of information retrieval based on a terminological logic.
In Korfhage, R., Rasmussen, E., and Willett, P., editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 298–308, New York. ACM.

Metzler, D. and Croft, W. B. (2004).
Combining the language model and inference network approaches to retrieval.
Information Processing & Management, 40(5):735–750.

Nie, J. (1992).
Towards a probabilistic modal logic for semantic-based information retrieval.
In Belkin, N., Ingwersen, P., and Pejtersen, M., editors, *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 140–151, New York.

Ponte, J. and Croft, W. (1998).
A language modeling approach to information retrieval.
In Croft, W. B., Moffat, A., van Rijsbergen, C. J., Wilkinson, R., and Zobel, J., editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, New York. ACM.

Robertson, S. (2004).
Understanding inverse document frequency: On theoretical arguments for idf.
Journal of Documentation, 60:503–520.

Robertson, S. (2005).
On event spaces and probabilistic models in information retrieval.
Information Retrieval Journal, 8(2):319–329.

Robertson, S. and Sparck Jones, K. (1976).
Relevance weighting of search terms.
Journal of the American Society for Information Science, 27:129–146.

Robertson, S. E. and Walker, S. (1994).

Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval.
In [Croft and van Rijsbergen, 1994], pages 232–241.

Robertson, S. E., Walker, S., and Hancock-Beaulieu, M. (1995).
Large test collection experiments on an operational interactive system: Okapi at TREC.
Information Processing and Management, 31:345–360.

Rocchio, J. (1966).
Document retrieval systems - optimization and evaluation.
Report ISR-10 to the NSF, Computation Laboratory, Harvard University.

Rocchio, J. (1971).
Relevance feedback in information retrieval.
In [Salton, 1971].

Roelleke, T., Tsikrika, T., and Kazai, G. (2006).
A general matrix framework for modelling information retrieval.
Journal on Information Processing & Management (IP&M), Special Issue on Theory in Information Retrieval, 42(1).

Roelleke, T. and Wang, J. (2006).
A parallel derivation of probabilistic information retrieval models.
In *ACM SIGIR*, pages 107–114, Seattle, USA.

Salton, G., editor (1971).
The SMART Retrieval System - Experiments in Automatic Document Processing.
Prentice Hall, Englewood, Cliffs, New Jersey.

Salton, G., Fox, E., and Wu, H. (1983).

Extended Boolean information retrieval.
Communications of the ACM, 26:1022–1036.

Salton, G., Wong, A., and Yang, C. (1975).
A vector space model for automatic indexing.
Communications of the ACM, 18:613–620.

Sebastiani, F. (2002).
Machine learning in automated text categorization.
ACM Comput. Surv., 34(1):1–47.

Turtle, H. and Croft, W. (1991a).
Efficient probabilistic inference for text retrieval.
In *Proceedings RIAO 91*, pages 644–661, Paris, France.

Turtle, H. and Croft, W. (1991b).
Evaluation of an inference network-based retrieval model.
ACM Transactions on Information Systems, 9(3):187–222.

Turtle, H. and Croft, W. (1992).
A comparison of text retrieval models.
The Computer Journal, 35.

Turtle, H. and Croft, W. B. (1990).
Inference networks for document retrieval.
In Vidick, J.-L., editor, *Proceedings of the 13th International Conference on Research and Development in Information Retrieval*, pages 1–24, New York. ACM.

van Rijsbergen, C. J. (1979).

Information Retrieval.
Butterworths, London, 2. edition.
http://www.dcs.glasgow.ac.uk/Keith/Preface.html.

van Rijsbergen, C. J. (1986).
A non-classical logic for information retrieval.
The Computer Journal, 29(6):481–485.

van Rijsbergen, C. J. (1989).
Towards an information logic.
In Belkin, N. and van Rijsbergen, C. J., editors, *Proceedings of the Twelfth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 77–86, New York.

Wong, S. and Yao, Y. (1995).
On modeling information retrieval with probabilistic inference.
ACM Transactions on Information Systems, 13(1):38–68.

Zaragoza, H., Hiemstra, D., and Tipping, M. (2003).
Bayesian extension to the language model for ad hoc information retrieval.
In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 4–9, New York, NY, USA. ACM Press.