Retrieval Models

406.424 Internet Applications

Jonghun Park

jonghun@snu.ac.kr

Dept. of Industrial Eng.

Seoul National University

9/1/2010

retrieval models

- theories about relevance
- provide a mathematical framework for defining the search process
  - includes explanation of assumptions
  - basis of many ranking algorithms
- progress in retrieval models has corresponded with improvements in effectiveness

relevance

- complex concept that has been studied for some time
  - many factors to consider
  - people often disagree when making relevance judgments
- retrieval models make various assumptions about relevance to simplify problem
  - e.g., topical vs. user relevance
  - e.g., binary vs. multi-valued relevance

retrieval model overview

- older models
  - boolean retrieval
  - vector space model
- probabilistic models
  - BM25
  - language models
- combining evidence
  - inference networks
  - learning to rank

boolean retrieval

- 2 possible outcomes for query processing
  - TRUE and FALSE
  - “exact-match” retrieval
  - simplest form of ranking
- query usually specified using boolean operators
  - AND, OR, NOT
  - proximity operators also used

boolean retrieval

- advantages
  - results are predictable, relatively easy to explain
  - many different features can be incorporated
  - efficient processing since many documents can be eliminated from search
- disadvantages
  - effectiveness depends entirely on user
  - simple queries usually don’t work well
  - complex queries are difficult

searching by numbers

- sequence of queries driven by number of retrieved documents
  - e.g. “lincoln” search of news articles
  - president AND lincoln
  - president AND lincoln AND NOT (automobile OR car)
  - president AND lincoln AND biography AND life AND birthplace AND gettysburg AND NOT (automobile OR car)
    - anything retrieved?
  - president AND lincoln AND (biography OR life OR birthplace OR gettysburg) AND NOT (automobile OR car)

vector space model

- documents and query represented by a **vector of term weights**
- collection represented by a matrix of term weights

$$Q = (q_1, q_2, \ldots, q_t)$$
$$D_i = (d_{i1}, d_{i2}, \ldots, d_{it})$$

	$Term_1$	$Term_2$	$\ldots$	$Term_t$
Doc_1	d_{11}	d_{12}	$\ldots$	d_{1t}
Doc_2	d_{21}	d_{22}	$\ldots$	d_{2t}
$\vdots$	$\vdots$			
Doc_n	d_{n1}	d_{n2}	$\ldots$	d_{nt}

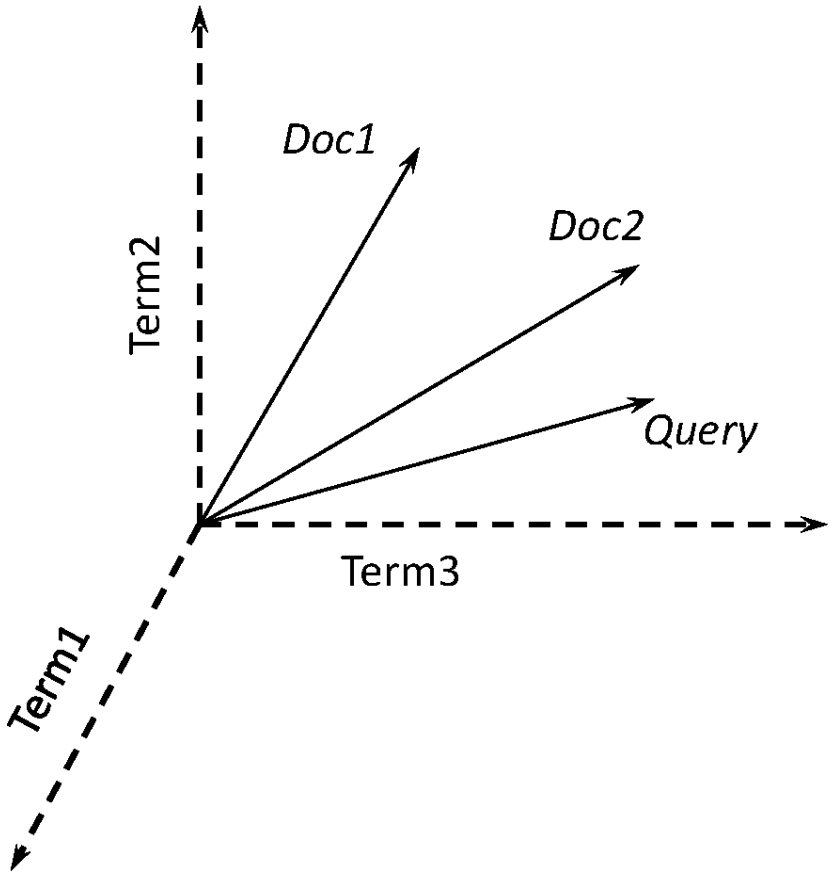
vector space model example

D_1 Tropical Freshwater Aquarium Fish.
D_2 Tropical Fish, Aquarium Care, Tank Setup.
D_3 Keeping Tropical Fish and Goldfish in Aquariums, and Fish Bowls.
D_4 The Tropical Tank Homepage - Tropical Fish and Aquariums.

Terms	Documents			
	D_1	D_2	D_3	D_4
aquarium	1	1	1	1
bowl	0	0	1	0
care	0	1	0	0
fish	1	1	2	1
freshwater	1	0	0	0
goldfish	0	0	1	0
homepage	0	0	0	1
keep	0	0	1	0
setup	0	1	0	0
tank	0	1	0	1
tropical	1	1	1	2

vector space model

- 3-d pictures useful, but can be misleading for high-dimensional space



vector space model

- documents ranked by distance between points representing query and documents
  - **similarity** measure more common than a distance or **dissimilarity** measure
  - e.g. cosine correlation

$$Cosine(D_i, Q) = \frac{\sum_{j=1}^{t} d_{ij} \cdot q_j}{\sqrt{\sum_{j=1}^{t} {d_{ij}}^2 \cdot \sum_{j=1}^{t} {q_j}^2}}$$

similarity calculation

- consider two documents D_1, D_2 and a query Q
  - D_1 = (0.5, 0.8, 0.3), D_2 = (0.9, 0.4, 0.2), Q = (1.5, 1.0, 0)

$$Cosine(D_1, Q) = \frac{(0.5 \times 1.5) + (0.8 \times 1.0)}{\sqrt{(0.5^2 + 0.8^2 + 0.3^2)(1.5^2 + 1.0^2)}}$$
$$= \frac{1.55}{\sqrt{(0.98 \times 3.25)}} = 0.87$$

$$Cosine(D_2, Q) = \frac{(0.9 \times 1.5) + (0.4 \times 1.0)}{\sqrt{(0.9^2 + 0.4^2 + 0.2^2)(1.5^2 + 1.0^2)}}$$
$$= \frac{1.75}{\sqrt{(1.01 \times 3.25)}} = 0.97$$

term weights

- *tf.idf* weight
  - term frequency weight measures importance in **document**:

$$tf_{ik} = \frac{f_{ik}}{\sum_{j=1}^{t} f_{ij}}$$

    - where f_{ik} is # of occurrences of term k in D_i
  - inverse document frequency (IDF) measures importance in **collection**:

$$idf_k = \log \frac{N}{n_k}$$

  - an empirical, heuristic modification

$$d_{ik} = \frac{(\log(f_{ik})+1)\cdot\log(N/n_k)}{\sqrt{\sum_{k=1}^{t}[(\log(f_{ik})+1.0)\cdot\log(N/n_k)]^2}}$$

relevance feedback

- Rocchio algorithm
- optimal query
  - maximizes the difference between the average vector representing the relevant documents and the average vector representing the non-relevant documents
- modifies the initial weights in query vector to

$$q'_j = \alpha . q_j + \beta . \frac{1}{|Rel|} \sum_{D_i \in Rel} d_{ij} - \gamma . \frac{1}{|Nonrel|} \sum_{D_i \in Nonrel} d_{ij}$$

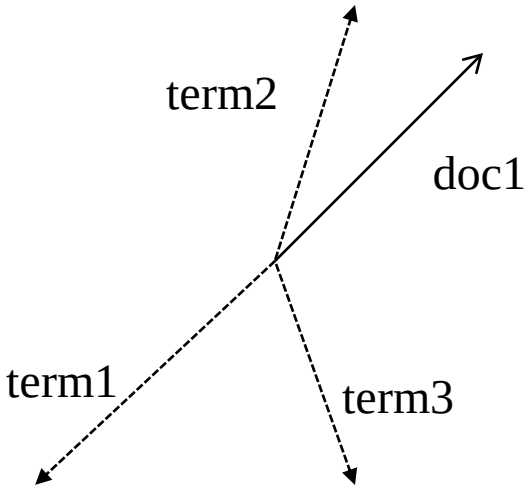
  - α, β, and γ are parameters (typical values: 8, 16, 4)
  - query terms with negative weights are dropped

vector space model

- advantages
  - simple computational framework for ranking
  - any similarity measure or term weighting scheme could be used
- disadvantages
  - assumption of term independence
  - no **predictions** about techniques for effective ranking

an extension of VSM (Park et al., IPL 2010)

- utilizes term-to-term similarity to compute the similarity between term vectors
- consistency of **term similarity matrix** C
  - iff a set of linearly independent vectors $\{t^1, t^2, \ldots, t^n\}$ s.t. $(t^i) \cdot (t^j)^T = \cos\theta_{i,j} = c_{i,j}, \forall i, j$ exist
- C is consistent iff C is positive definite



$$C' = L'(L')^T = \begin{pmatrix} 1 & 0.1 & 0.2 \\ 0.1 & 1 & 0.1 \\ 0.2 & 0.1 & 1 \end{pmatrix}$$

$$\alpha = (1, 6, 2, 3) \rightarrow \alpha' = (1, 6 + 2, 3)$$

$$\beta = (2, 1, 1, 7) \rightarrow \beta' = (2, 1 + 1, 7)$$

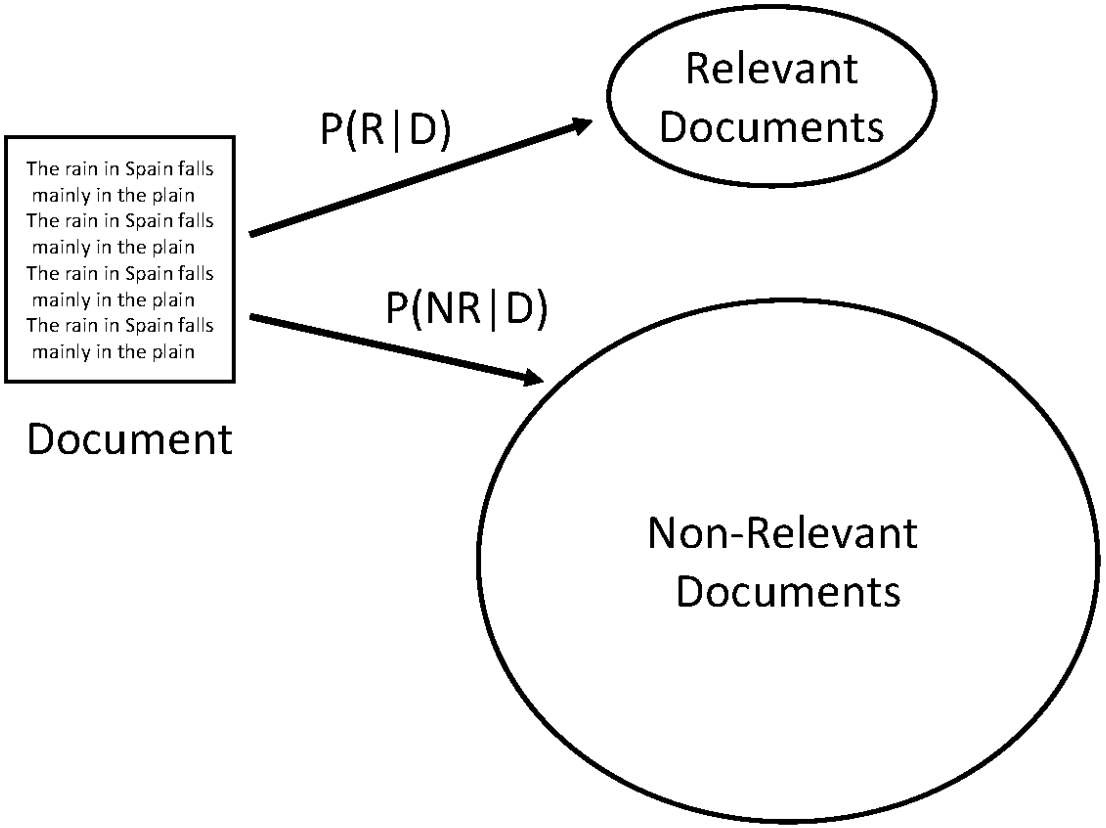
$$\cos\theta_{\alpha',\beta'} = \frac{\alpha' C' (\beta')^T}{\sqrt{\alpha' C' (\alpha')^T \beta' C' (\beta')^T}} \approx 0.675$$

probability ranking principle

- Robertson (1977)
  - "If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of **decreasing probability of relevance** to the user who submitted the request,
  - where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose,
  - the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data."
- doesn't tell us how to calculate the probability of relevance

IR as classification

- $P(R|D)$: probability of relevance given the representation of document D
- $P(NR|D)$: probability of non-relevance given the representation of document D



Bayes classifier

- Bayes decision rule
  - a document D is relevant if $P(R|D) > P(NR|D)$
- estimating probabilities
  - use Bayes rule

    $$P(R|D) = \frac{P(D|R)P(R)}{P(D)}$$
  - classify a document as relevant if

    $$P(R|D) > P(NR|D)$$

    $$\Rightarrow \frac{P(D|R)P(R)}{P(D)} > \frac{P(D|NR)P(NR)}{P(D)}$$

    $$\Rightarrow \frac{P(D|R)}{P(D|NR)} > \frac{P(NR)}{P(R)}$$

estimating $P(D|R)$

- interpretations
  - $P(R)$: prob. that a randomly chosen document is relevant
  - $P(D|R)$: prob. that if a relevant document is retrieved, then that document's representation is D
- assume term independence
  - $P(D|R) = \prod_{i=1}^{t} P(d_i|R)$
  - $P(d_i|R)$: prob. that a relevant document will contain term i
  - where $D = \{d_1, d_2, \ldots, d_t\}$, $d_i = 1$ if term i is present in D; 0, o.w.
- example
  - $D = (1,0,1)$
  - $P(D|R) = p_1 \times (1\text{-}p_2) \times p_3$
    where p_i is the probability that term i occurs in a document from the relevant set

binary independence model

- document represented by a vector of **binary features** indicating term occurrence (or non-occurrence)
- p_i is probability that term *i* occurs (i.e., has value 1) in relevant document
- s_i is probability that term *i* occurs in non-relevant document

$$\frac{P(D|R)}{P(D|NR)} = \prod_{i:d_i=1} \frac{p_i}{s_i} \cdot \prod_{i:d_i=0} \frac{1-p_i}{1-s_i}$$

$$= \prod_{i:d_i=1} \frac{p_i}{s_i} \cdot (\prod_{i:d_i=1} \frac{1-s_i}{1-p_i} \cdot \prod_{i:d_i=1} \frac{1-p_i}{1-s_i}) \cdot \prod_{i:d_i=0} \frac{1-p_i}{1-s_i}$$

$$= \prod_{i:d_i=1} \frac{p_i(1-s_i)}{s_i(1-p_i)} \cdot \prod_i \frac{1-p_i}{1-s_i}$$

the second product is same for all documents!

binary independence Model

- scoring function is

$$\sum_{i:d_i=1} \log \frac{p_i(1-s_i)}{s_i(1-p_i)}$$

- query provides information about relevant documents
  - for the term *i* not in query, $p_i = s_i$
- if we have no other information about the relevant set
  - we may assume $p_i = 0.5$
  - s_i is estimated by using the TF in the whole collection (assuming that # of relevant documents is much smaller than the total # of documents)

$$\log \frac{0.5(1-\frac{n_i}{N})}{\frac{n_i}{N}(1-0.5)} = \log \frac{N-n_i}{n_i}$$

contingency table

- when we have information about term occurrences in the relevant and non-relevant sets,
  - r_i: # of relevant documents containing term i
  - n_i: # of documents containing term i
  - N: total # of documents
  - R: # of relevant documents for a query
- $p_i = r_i/R$ and $s_i = (n_i - r_i) / (N - R)$
- problem of **log0** in the scoring function

	Relevant	Non-relevant	Total
$d_i = 1$	r_i	$n_i - r_i$	n_i
$d_i = 0$	$R - r_i$	$N - n_i - R + r_i$	$N - r_i$
Total	R	$N - R$	N

contingency table

- better estimates would be

$$p_i = (r_i + 0.5)/(R + 1)$$
$$s_i = (n_i - r_i + 0.5)/(N - R + 1)$$

(note) without any relevance information, $r_i = R = 0 \Rightarrow p_i = 0.5$

- gives a scoring function

$$\sum_{i:d_i=q_i=1} \log \frac{(r_i+0.5)/(R-r_i+0.5)}{(n_i-r_i+0.5)/(N-n_i-R+r_i+0.5)}$$

- limitations: r_i for a query is not available in most cases, which makes the scoring function behave like *idf* term weighting

BM25

- popular and effective ranking algorithm based on **binary independence** model
  - incorporates document and query **TF weights**
  - not a formal model
  - r and R are set to 0 if there is no relevance information
  - f_i is TF of term i in the document
  - qf_i is TF of term i in the query
  - k_1, k_2 and K are parameters whose values are set empirically
  - dl is document length
  - typical TREC value for k_1 is 1.2, k_2 varies from 0 to 1000, b = 0.75

$$\sum_{i \in Q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(k_1+1) f_i}{K + f_i} \cdot \frac{(k_2+1) qf_i}{k_2 + qf_i}$$

$$K = k_1((1-b) + b \cdot \frac{dl}{avdl})$$

BM25 example

- query with two terms, "president lincoln", ($qf_1 = qf_2 = 1$)
- no relevance information (r and R are zero)
- N = 500,000 documents
- "president" occurs in 40,000 documents (n_1 = 40, 000)
- "lincoln" occurs in 300 documents (n_2 = 300)
- "president" occurs 15 times in doc (f_1 = 15)
- "lincoln" occurs 25 times (f_2 = 25)
- document length is 90% of the average length ($dl/avdl$ = 0.9)
- k_1 = 1.2, b = 0.75, and k_2 = 100
- $K = 1.2 \cdot (0.25 + 0.75 \cdot 0.9) = 1.11$

BM25 Example

$$
\begin{aligned}
BM25(Q,D) \quad = \quad & \\
& \log \frac{(0+0.5)/(0-0+0.5)}{(40000-0+0.5)/(500000-40000-0+0+0.5)} \\
& \times \frac{(1.2+1)15}{1.11+15} \times \frac{(100+1)1}{100+1} \\
& + \log \frac{(0+0.5)/(0-0+0.5)}{(300-0+0.5)/(500000-300-0+0+0.5)} \\
& \times \frac{(1.2+1)25}{1.11+25} \times \frac{(100+1)1}{100+1}
\end{aligned}
$$

$$
\begin{aligned}
= \quad & \log 460000.5/40000.5 \cdot 33/16.11 \cdot 101/101 \\
& + \log 499700.5/300.5 \cdot 55/26.11 \cdot 101/101 \\
= \quad & 2.44 \cdot 2.05 \cdot 1 + 7.42 \cdot 2.11 \cdot 1 \\
= \quad & 5.00 + 15.66 = 20.66
\end{aligned}
$$

BM25 example

- effect of term frequencies

Frequency of "president"	Frequency of "lincoln"	BM25 score
15	25	20.66
15	1	12.74
15	0	5.00
1	25	18.2
0	25	15.66

language model

- unigram language model
  - probability distribution over the words in a language
  - generation of text consists of pulling words out of a “bucket” according to the probability distribution and replacing them
  - previous words have no impact on the prediction of a next word
- n-gram language model
  - some applications use bigram and trigram language models where probabilities depend on previous words
  - an n-gram model predicts a word based on the previous *n*-1 words

language model

- a topic in a document or query can be represented as a language model
  - define a **topic** as a **probability distribution over words**
  - i.e., words that tend to occur often when discussing a topic will have high probabilities in the corresponding language model
- **multinomial** distribution over words
  - text is modeled as a finite sequence of words, where there are *t* possible words at each point in the sequence
  - commonly used, but not only possibility
  - doesn't model burstiness, which is the observation that once a word is "pulled out of the bucket", it tends to be pulled out repeatedly

LMs for retrieval

- 3 possibilities:
  - probability of generating the query text from a document language model
  - probability of generating the document text from a query language model
  - comparing the language models representing the query and document topics
- models of topical relevance

query likelihood model

- rank documents by the probability that the query could be generated by the document model (i.e. same topic)
- given query, start with $P(D|Q)$
  - query likelihood given the document
- using Bayes' rule

$$p(D|Q) \stackrel{rank}{=} P(Q|D)P(D)$$

- assuming $P(D)$ is uniform, unigram model ranks documents by

$$P(Q|D) = \prod_{i=1}^{n} P(q_i|D)$$

estimating probabilities

- obvious estimate for unigram probabilities is based on the following **maximum likelihood estimate** that makes the observed value of $f_{q;D}$ most likely:

$$P(q_i|D) = \frac{f_{q_i,D}}{|D|}$$

  - $f_{q;D}$: # of times word q_i occurs in document D
  - $|D|$: # of words in D

- if query words are missing from document, score will be zero
  - missing 1 out of 4 query words same as missing 3 out of 4

smoothing

- document texts are a **sample** from the language model
  - Missing words should not have zero probability of occurring
- smoothing is a technique for estimating probabilities for missing (or unseen) words
  - lower (or discount) the probability estimates for words that are seen in the document text
  - assign that “left-over” probability to the estimates for the words that are not seen in the text

estimating probabilities

- estimate for unseen words is $\alpha_D P(q_i|C)$
  - $P(q_i|C)$ is the probability for query word i in the **collection** language model for collection C (background probability)
  - α_D is a parameter
- estimate for words that occur is

$$(1 - \alpha_D)\, P(q_i|D) + \alpha_D\, P(q_i|C)$$

- different forms of estimation come from different α_D

Jelinek-Mercer smoothing

- α_D is a constant, λ
- gives estimate of

$$p(q_i|D) = (1-\lambda)\frac{f_{q_i,D}}{|D|} + \lambda\frac{c_{q_i}}{|C|}$$

  - c_{q_i} is # of times of a query word occurs in the collection of docs
  - $|C|$: total # of word occurrences in the collection
- ranking score

$$P(Q|D) = \prod_{i=1}^{n}((1-\lambda)\frac{f_{q_i,D}}{|D|} + \lambda\frac{c_{q_i}}{|C|})$$

- use logs for convenience
  - to avoid accuracy problems multiplying small numbers

$$\log P(Q|D) = \sum_{i=1}^{n}\log((1-\lambda)\frac{f_{q_i,D}}{|D|} + \lambda\frac{c_{q_i}}{|C|})$$

where is *tf.idf* weight?

$$\log P(Q|D) = \sum_{i=1}^{n} \log((1-\lambda)\frac{f_{q_i,D}}{|D|} + \lambda\frac{c_{q_i}}{|C|})$$

$$= \sum_{i:f_{q_i,D}>0} \log((1-\lambda)\frac{f_{q_i,D}}{|D|} + \lambda\frac{c_{q_i}}{|C|}) + \sum_{i:f_{q_i,D}=0} \log(\lambda\frac{c_{q_i}}{|C|})$$

$$= \sum_{i:f_{q_i,D}>0} \log\frac{((1-\lambda)\frac{f_{q_i,D}}{|D|} + \lambda\frac{c_{q_i}}{|C|})}{\lambda\frac{c_{q_i}}{|C|}} + \sum_{i=1}^{n} \log(\lambda\frac{c_{q_i}}{|C|})$$

$$\overset{rank}{=} \sum_{i:f_{q_i,D}>0} \log\left(\frac{((1-\lambda)\frac{f_{q_i,D}}{|D|}}{\lambda\frac{c_{q_i}}{|C|}} + 1\right)$$

- directly proportional to the document TF
- inversely proportional to the collection frequency (i.e. DF)

Dirichlet smoothing

- α_D depends on document length

  $$\alpha_D = \frac{\mu}{|D|+\mu}$$

  - μ is a parameter whose value is set empirically
- gives probability estimation of

  $$p(q_i|D) = \frac{f_{q_i,D}+\mu\frac{c_{q_i}}{|C|}}{|D|+\mu}$$
- and document score

  $$\log P(Q|D) = \sum_{i=1}^{n} \log \frac{f_{q_i,D}+\mu\frac{c_{q_i}}{|C|}}{|D|+\mu}$$
- generally more effective than Jelinek-Mercer, especially for the short queries

query likelihood example

- for the term “president”
  - $f_{qi,D} = 15$, $c_{qi} = 160{,}000$
- for the term “lincoln”
  - $f_{qi,D} = 25$, $c_{qi} = 2{,}400$
- number of word occurrences in the document $|D|$ is assumed to be 1,800
- number of word occurrences in the collection is 10^9
  - 500,000 documents times an average of 2,000 words
- $\mu = 2{,}000$

query likelihood example

$$
\begin{aligned}
QL(Q,D) &= \log \frac{15 + 2000 \times (1.6 \times 10^5 / 10^9)}{1800 + 2000} \\
&\quad + \log \frac{25 + 2000 \times (2400/10^9)}{1800 + 2000} \\
&= \log(15.32/3800) + \log(25.005/3800) \\
&= -5.51 + -5.02 = -10.53
\end{aligned}
$$

Frequency of "president"	Frequency of "lincoln"	QL score
15	25	-10.53
15	1	-13.75
15	0	-19.05
1	25	-12.99
0	25	-14.40

- negative number because summing logs of small numbers

relevance models

- relevance model
  - representation of the topic of a query as a language model
  - query and relevant documents are samples from this model
- $P(D|R)$: probability of generating the text in a document given a relevance model
  - **document likelihood** model: given some examples of relevant documents for a query, estimate the probabilities in the relevance model and use this model to predict the relevance of new documents
  - less effective than query likelihood due to difficulties comparing across documents of different lengths

pseudo-relevance feedback

- estimate relevance model from query and top-ranked documents
- rank documents by similarity of **document model** to **relevance model**
- **Kullback-Leibler divergence** (KL-divergence) is a well-known measure of the difference between two probability distributions

KL-divergence

- given the **true** probability distribution *P* and another distribution *Q* that is an **approximation** to *P*,

$$KL(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- use **negative** KL-divergence for ranking and assume the true distribution to be the relevance model for the query (*R*) and the approximation to be the document language model (*D*) is the true distribution (not symmetric)
  - *V* is the set of vocabulary

$$\sum_{w \in V} P(w|R) \log P(w|D) - \sum_{w \in V} P(w|R) \log P(w|R)$$

(note) the second term does not depend on the document, and can be ignored for ranking

KL-Divergence

- given a simple maximum likelihood estimate for $P(w|R)$, based on the frequency in the query text ($f_{w,Q}$) and the # of words in the query ($|Q|$), ranking score is

$$\sum_{w \in V} \frac{f_{w,Q}}{|Q|} \log P(w|D)$$

  - rank-equivalent to query likelihood score
- query likelihood model is a special case of retrieval based on relevance model

estimating the relevance model

- probability of pulling a word *w* out of the "bucket" representing the relevance model depends on the *n* query words we have just pulled out

$$P(w|R) \approx P(w|q_1 \ldots q_n)$$

- by definition

$$P(w|R) \approx \frac{P(w,q_1 \ldots q_n)}{P(q_1 \ldots q_n)}$$

estimating the relevance model

- joint probability is

$$P(w, q_1 \dots q_n) = \sum_{D \in \mathcal{C}} p(D) P(w, q_1 \dots q_n | D)$$

- assume the independence

$$P(w, q_1 \dots q_n | D) = P(w|D) \prod_{i=1}^{n} P(q_i|D)$$

- gives

$$P(w, q_1 \dots q_n) = \sum_{D \in \mathcal{C}} P(D) P(w|D) \prod_{i=1}^{n} P(q_i|D)$$

estimating the relevance model

- $P(D)$ usually assumed to be uniform and can be ignored
- $P(w, q_1 \ldots q_n)$ is simply a weighted average of the language model probabilities for w in a set of documents, where the weights are the **query likelihood scores** for those documents
- gives a formal model for pseudo-relevance feedback
  - query expansion technique

pseudo-relevance feedback algorithm

1. Rank documents using the query likelihood score for query Q.
2. Select some number of the top-ranked documents to be the set $\mathcal{C}$.
3. Calculate the relevance model probabilities $P(w|R)$. $P(q_1 \dots q_n)$ is used as a normalizing constant and is calculated as

$$P(q_1 \dots q_n) = \sum_{w \in V} P(w, q_1 \dots q_n)$$

4. Rank documents again using the KL-divergence score

$$\sum_{w} P(w|R) \log P(w|D)$$

- the document language model probability, $P(w|D)$, should be estimated using Dirichlet smoothing
- the summation in step 4 is typically done over a small number of the highest-probability words

example from top 10 docs

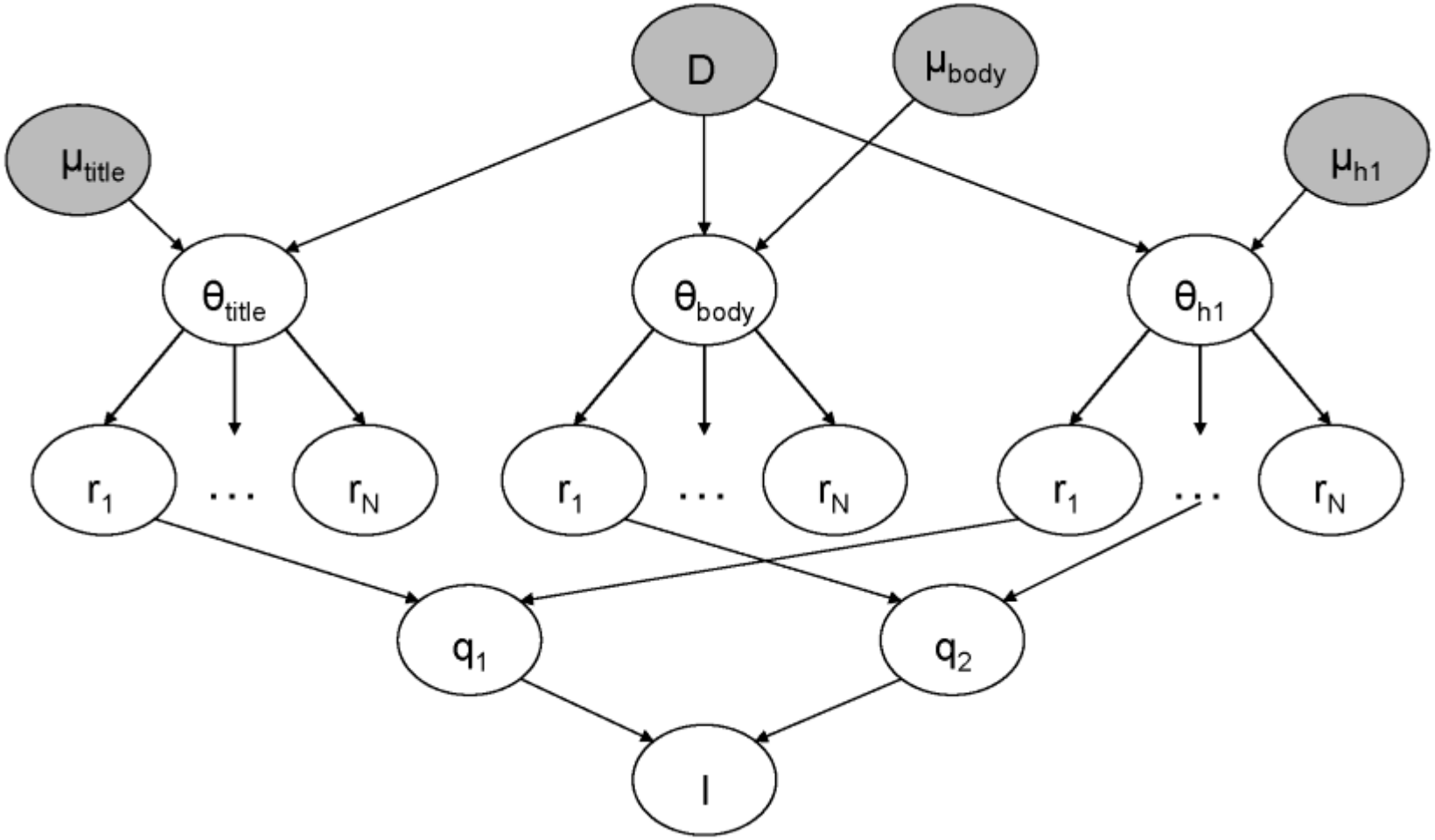
- highest-probability terms from relevance model (estimated using top 10 documents)

president lincoln	*abraham lincoln*	*fishing*	*tropical fish*
lincoln	lincoln	fish	fish
president	america	farm	tropic
room	president	salmon	japan
bedroom	faith	new	aquarium
house	guest	wild	water
white	abraham	water	species
america	new	caught	aquatic
guest	room	catch	fair
serve	christian	tag	china
bed	history	time	coral
washington	public	eat	source
old	bedroom	raise	tank
office	war	city	reef
war	politics	people	animal
long	old	fishermen	tarpon
abraham	national	boat	fishery

combining evidence

- effective retrieval requires the combination of many pieces of evidence about a document's potential relevance
  - have focused on simple word-based evidence
  - many other types of evidence
    - structure, PageRank, metadata, even scores from different models
- inference network model is one approach to combining evidence
  - uses Bayesian network formalism

inference network



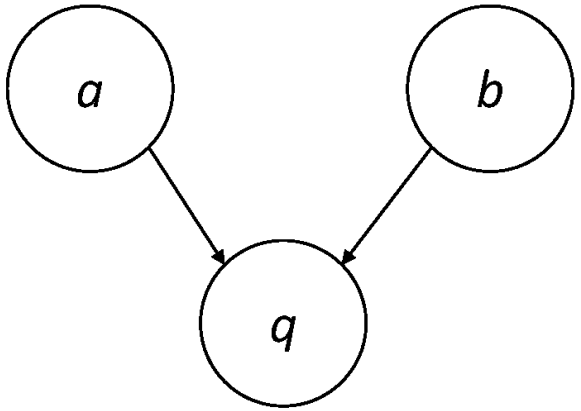
inference network

- **document node** (D) corresponds to the event that a document is observed
- **representation nodes** (r_i) are document features (evidence)
  - probabilities associated with those features are based on language models θ estimated using the parameters μ
  - one language model for each significant document structure
  - r_i nodes can represent proximity features, or other types of evidence (e.g. date)

inference network

- **query nodes** (q_i) are used to combine evidence from representation nodes and other query nodes
  - represent the occurrence of more complex evidence and document features
  - a number of combination operators are available
- **information need node** (I) is a special query node that combines all of the evidence from the other query nodes
  - network computes $P(I|D, \mu)$

example: AND combination



a and *b* are *parent* nodes for q

$P(q = \text{TRUE}\|a, b)$	a	b
0	FALSE	FALSE
0	FALSE	TRUE
0	TRUE	FALSE
1	TRUE	TRUE

example: AND combination

- combination must consider all possible states of parents
- let
  - $p_{i,j}$: prob that q is TRUE given that parents' states i and j
  - $bel_{and}(q)$: belief value (probability) that results from AND combination

$$
\begin{aligned}
bel_{and}(q) &= p_{00}P(a = \mathsf{FALSE})P(b = \mathsf{FALSE}) \\
&\quad + p_{01}P(a = \mathsf{FALSE})P(b = \mathsf{TRUE}) \\
&\quad + p_{10}P(a = \mathsf{TRUE})P(b = \mathsf{FALSE}) \\
&\quad + p_{11}P(a = \mathsf{TRUE})P(b = \mathsf{TRUE}) \\
&= 0 \cdot (1 - p_a)(1 - p_b) + 0 \cdot (1 - p_a)p_b + 0 \cdot p_a(1 - p_b) + 1 \cdot p_a p_b \\
&= p_a p_b
\end{aligned}
$$

inference network operators

$$
\begin{aligned}
bel_{not}(q) &= 1 - p_1 \\
bel_{or}(q) &= 1 - \prod_i^n (1 - p_i) \\
bel_{and}(q) &= \prod_i^n p_i \\
bel_{wand}(q) &= \prod_i^n p_i^{wt_i} \\
bel_{max}(q) &= max\{p_1, p_2, \ldots, p_n\} \\
bel_{sum}(q) &= \frac{\sum_i^n p_i}{n} \\
bel_{wsum}(q) &= \frac{\sum_i^n wt_i p_i}{\sum_i^n wt_i}
\end{aligned}
$$

Galago query language

- a document is viewed as a sequence of text that may contain arbitrary tags
- for each tag type T within a document (e.g., title, body, h1, …), we define the **context** of T to be all of the text and tags that appear within tags of type T

```
<html>
<head>
<title>Department Descriptions</title>
</head>
<body>
The following list describes ...
<h1>Agriculture</h1> ...
<h1>Chemistry</h1> ...
<h1>Computer Science</h1> ...
<h1>Electrical Engineering</h1> ...
</body>
</html>
```

title context:

```
<title>Department Descriptions</title>
```

h1 context:

```
<h1>Agriculture</h1>
<h1>Chemistry</h1> ...
<h1>Computer Science</h1> ...
<h1>Electrical Engineering</h1> ...
```

body context:

```
<body> The following list describes ...
<h1>Agriculture</h1> ...
<h1>Chemistry</h1> ...
<h1>Computer Science</h1> ...
<h1>Electrical Engineering</h1> ...
</body>
```


Galago query language examples

Examples:

#combine(#syn(dog canine) training) – rank by two terms, one of which is a synonym.

#combine(biography #syn(#od:1(president lincoln) #od:1(abraham lincoln))) – rank using two terms, one of which is a synonym of "president lincoln" and "abraham lincoln".

#weight(1.0 #od:1(civil war) 3.0 lincoln 2.0 speech) – rank using three terms, and weight the term "lincoln" as most important, followed by "speech", then "civil war".

#filter(aquarium #combine(tropical fish)) – consider only those documents containing the word "aquarium" and "tropical" or "fish", and rank them according to the query #combine(aquarium #combine(tropical fish)).

#filter(#od:1(john smith).author) #weight(2.0 europe 1.0 travel) – rank documents about "europe" or "travel" that have "John Smith" in the author context.

- (note) terms correspond to representation nodes in the inference network model

web search

- major differences to TREC news
  - size of collection
  - connections between documents
  - range of document types
  - importance of spam
  - volume of queries
  - range of query types

search taxonomy

- informational
  - finding information about some topic which may be on one or more web pages
  - topical search
- navigational
  - finding a particular web page that the user has either seen before or is assumed to exist
- transactional
  - finding a site where a task such as shopping or downloading music can be performed

web search

- for effective navigational and transactional search, need to combine features that reflect **user relevance**
- commercial web search engines combine evidence from **hundreds of features** to generate a ranking score for a web page
  - page content, page metadata, anchor text, links (e.g., PageRank), and user behavior (click logs)
  - page metadata: e.g., “age”, how often it is updated, the URL of the page, the domain name of its site, and the amount of text content
- much of the evidence that is crucial for effective navigational search is not important for topical searches

search engine optimization (SEO)

- understanding the relative importance of features used in search and how they can be manipulated to obtain better search rankings for a web page
  - e.g., improve the text used in the title tag, improve the text in heading tags, make sure that the domain name and URL contain important keywords, and try to improve the anchor text and link structure
  - some of these techniques are regarded as not appropriate by search engine companies

web search

- in TREC evaluations, most effective features for navigational search are:
  - text in the title, body, and heading (h1, h2, h3, and h4) parts of the document, the anchor text of all links pointing to the document, the PageRank number, and the inlink count
- given size of web, **many pages will contain all query terms**
  - ranking algorithm focuses on discriminating between these pages
  - **word proximity** is important

term proximity

- many models have been developed
- **n-grams** are commonly used in commercial web search
- **dependence model** based on the assumption that query terms are likely to appear in close proximity to each other within relevant documents has been effective
  - e.g. query = “embryonic stem cells”

```
#weight(
        0.8 #combine(embryonic stem cells)
        0.1 #combine( #od:1(stem cells) #od:1(embryonic stem)
                      #od:1(embryonic stem cells))
        0.1 #combine( #uw:8(stem cells) #uw:8(embryonic cells)
                      #uw:8(embryonic stem) #uw:12(embryonic stem cells)))
```

example web query

- a complex Galago query expression can be generated from a simple user query

```
#weight(
        0.1 #weight( 0.6 #prior(pagerank) 0.4 #prior(inlinks))
        1.0 #weight(
            0.9 #combine(
                #weight( 1.0 pet.(anchor) 1.0 pet.(title)
                         3.0 pet.(body) 1.0 pet.(heading))
                #weight( 1.0 therapy.(anchor) 1.0 therapy.(title)
                         3.0 therapy.(body) 1.0 therapy.(heading)))
            0.1 #weight(
                1.0 #od:1(pet therapy).(anchor) 1.0 #od:1(pet therapy).(title)
                3.0 #od:1(pet therapy).(body) 1.0 #od:1(pet therapy).(heading))
            0.1 #weight(
                1.0 #uw:8(pet therapy).(anchor) 1.0 #uw:8(pet therapy).(title)
                3.0 #uw:8(pet therapy).(body) 1.0 #uw:8(pet therapy).(heading)))
        )
```


machine learning and IR

- considerable interaction between these fields
  - Rocchio algorithm (60s) is a simple learning approach
  - 80s, 90s: learning ranking algorithms based on user feedback
  - 2000s: text categorization
- limited by amount of training data
- web query logs have generated new wave of research
  - e.g., “learning to rank”

generative vs. discriminative models

- all of the probabilistic retrieval models presented so far fall into the category of **generative** models
  - a generative model assumes that documents were generated from some underlying model (in this case, usually a multinomial distribution) and uses training data to estimate the parameters of the model
  - probability of belonging to a class (i.e. the relevant documents for a query) is then estimated using Bayes' Rule and the document model

generative vs. discriminative models

- a **discriminative** model estimates the probability of belonging to a class directly from the observed features of the document based on the training data
- generative models perform well with low numbers of training examples
- discriminative models usually have the advantage given **enough training data**
  - can also easily incorporate many features

discriminative models for IR

- discriminative models can be trained using **explicit relevance judgments** or **click data** in query logs
  - click data is much cheaper, more noisy
- Ranking SVM (Support Vector Machine) takes as input **partial rank** information for queries
  - partial information about which documents should be ranked higher than others

ranking SVM

- training data is

$$(q_1, r_1), (q_2, r_2), \ldots, (q_n, r_n)$$

  - r is partial information about the desired ranking
    - if document d_a should be ranked higher than d_b, then $(d_a, d_b) \in r_i$
  - partial rank information comes from relevance judgments (allows multiple levels of relevance) or click data
    - e.g., d_1, d_2 and d_3 are the documents in the first, second and third rank of the search output, only d_3 clicked on → (d_3, d_1) and (d_3, d_2) will be in desired ranking for this query

ranking SVM

- learning a linear ranking function $\vec{w}.\vec{d}_a$
  - where w is a weight vector that is adjusted by learning
  - d_a is the vector representation of the features of document
  - non-linear functions also possible
- weights represent importance of features
  - learned using training data
  - e.g.,

$$\vec{w}.\vec{d} = (2, 1, 2).(2, 4, 1) = 2.2 + 1.4 + 2.1 = 10$$

ranking SVM

- learn *w* that satisfies as many of the following conditions as possible:

$$\forall(d_i, d_j) \in r_1 \quad : \quad \vec{w}.\vec{d_i} > \vec{w}.\vec{d_j}$$

$$\ldots$$

$$\forall(d_i, d_j) \in r_n \quad : \quad \vec{w}.\vec{d_i} > \vec{w}.\vec{d_j}$$

- can be formulated as an **optimization** problem

ranking SVM

$$
\begin{aligned}
minimize: \quad & \frac{1}{2}\vec{w}.\vec{w} + C\sum \xi_{i,j,k} \\
subject\ to: & \\
\forall (d_i, d_j) \in r_1 \quad : \quad & \vec{w}.\vec{d_i} > \vec{w}.\vec{d_j} + 1 - \xi_{i,j,1} \\
& \ldots \\
\forall (d_i, d_j) \in r_n \quad : \quad & \vec{w}.\vec{d_i} > \vec{w}.\vec{d_j} + 1 - \xi_{i,j,n} \\
& \forall i \forall j \forall k : \xi_i, j, k \geq 0
\end{aligned}
$$

- ξ, known as a slack variable, allows for misclassification of difficult or noisy training examples, and C is a parameter that is used to prevent overfitting

ranking SVM

- software available to do optimization: e.g., SVM^{light}
- each pair of documents in our training data can be represented by the vector:

  $(\vec{d}_i - \vec{d}_j)$
- score for this pair is:

  $\vec{w}.(\vec{d}_i - \vec{d}_j)$
- SVM classifier will find a classifier (i.e., $\vec{w}$) that makes the smallest score as large as possible
  - make the differences in scores as large as possible for the pairs of documents that are hardest to rank

topic models

- vocabulary mismatch problem
  - relevant documents do not match a query because they are using **different words** to describe the **same topic**
- improved representations of documents
  - can also be viewed as improved smoothing techniques
  - improve estimates for words that are related to the topic(s) of the document
- approaches
  - Latent Semantic Indexing (LSI)
  - Probabilistic Latent Semantic Indexing (pLSI)
  - Latent Dirichlet Allocation (LDA)

LDA

- model document as being generated from a **mixture** of topics
  - each topic is a language model
- LDA process for generating a document:

1. For each document D, pick a multinomial distribution θ_D from a Dirichlet distribution with parameter α ,
2. For each word position in document D,
   (a) pick a topic z from the multinomial distribution θ_D ,
   (b) Choose a word w from $P(w|z, \beta)$, a multinomial probability conditioned on the topic z with parameter β.

LDA

- a variety of techniques are available for learning the topic models and θ distributions using the document collection as the training data (but slow)
- given these distributions, language model probabilities for the words in documents are

$$P_{lda}(w|D) = P(w|\theta_D, \beta) = \sum_z P(w|z, \beta)P(z|\theta_D)$$

- used to **smooth the document representation** by mixing them with the query likelihood probability as follows:

$$P(w|D) = \lambda \left(\frac{f_{w,D} + \mu \frac{c_w}{|C|}}{|D| + \mu} \right) + (1 - \lambda) P_{lda}(w|D)$$

LDA

- that is, the mixture of the maximum likelihood probabilities, collection probabilities, and the LDA probabilities
- if the LDA probabilities are used directly as the document representation, the effectiveness will be significantly reduced because the features are too smoothed
  - e.g., in typical TREC experiment, only 400 topics used for the entire collection
  - generating LDA topics is expensive
- when used for smoothing, effectiveness is improved

LDA example

- highest-probability terms from 4 LDA topics (from TREC news)

Arts	*Budgets*	*Children*	*Education*
new	million	children	school
film	tax	women	students
show	program	people	schools
music	budget	child	education
movie	billion	years	teachers
play	federal	families	high
musical	year	work	public
best	spending	parents	teacher
actor	new	says	bennett
first	state	family	manigat
york	plan	welfare	namphy
opera	money	men	state
theater	programs	percent	president
actress	government	care	elementary
love	congress	life	haiti

application-based ranking models

- construct a test collection of queries, documents, and relevance judgments so that different versions of the ranking algorithm can be compared
- identify what evidence or features might be used to represent documents
  - terms and proximity terms are almost always useful
  - application-specific thesaurus can make a significant difference to ranking effectiveness
- decide how to combine the features to calculate a document score
  - open source search engines can be used
  - much of the time will be spent on tuning the retrieval effectiveness

best retrieval model depends on application and data available!