

8. Probabilistic Retrieval

Outline

- ① BIR: Binary Independence Retrieval
- ② The Ideal Answer
- ③ Formal Definition
- ④ Relevance Feedback
- ⑤ BM25

What should be calculated?

- So far: no theoretically based models, but still good results
- Goal: theoretical foundation for retrieval
- Probability that a document D is relevant with respect to a query Q
- Chance:
 - ▶ probability for relevance/probability for irrelevance

8.1 BIR: Binary Independence Retrieval

- Robertson/Sparck Jones [1976]
- Basic assumptions:
 - ▶ binary term weights, i.e. $w_{dk} \in \{0, 1\}$
 - ▶ cf. $[0, 1]$ for vector space model
 - ▶ independence assumption of the individual dimensions of the representation vector
 - ▶ same as in the vector space model

Definitions

- Document: D
- Query: Q
- Set of relevant documents: $R^+(Q)$
- Probability that D is relevant to Q

$$P(D \in R^+(Q))$$

- Representation of a document: $\beta(D)$
- Dimensionality (#terms): t

Core idea

- Conditional probability that document D' , which has the same representation as D , is relevant for query Q

$$P(D' \in R^+(Q) | \beta(D') = \beta(D))$$

- actually: probability of relevance for *all* documents of the same representation

Example

	D1	D2	D3	D4	D5	D6	D7	D8	D9
w_{d1}	1	1	1	0	0	0	1	1	0
w_{d2}	1	0	1	0	1	0	1	0	0
Relevance	✓	✓	✓	✗	✗	✗	✗	✗	✗

$$P(D' \in R^+(Q) | \beta(D') = x)$$

x	P(...)
(1,1)	$\frac{2}{3}$, since (1, 1) occurs 3x and 2x is relevant
(1,0)	$\frac{1}{2}$
...	...

Example - A new document

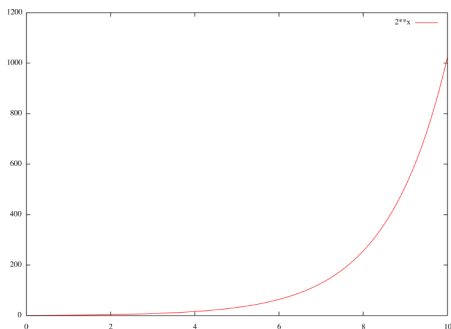
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
w_{d1}	1	1	1	0	0	0	1	1	0	1
w_{d2}	1	0	1	0	1	0	1	0	0	0
Relevance	✓	✓	✓	✗	✗	✗	✗	✗	✗	?

$$P(D' \in R^+(Q) | \beta(D') = x)$$

Relevance probability for $D10$ is $\frac{1}{2}$, because it is in class $(1, 0)$

Open problem

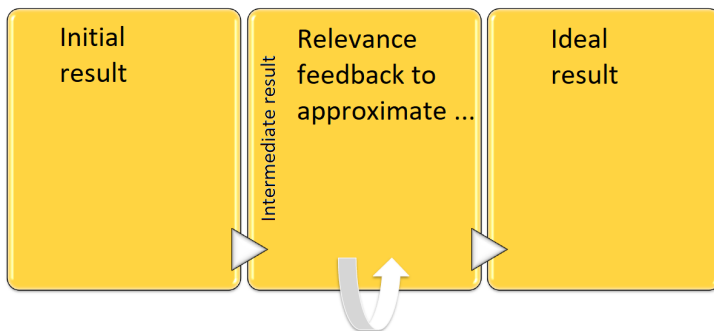
- How is the relevance probability determined empirically?
- 2^t possible representations
- Much more than 2^t documents are needed for evaluation!



8.2 The Ideal Answer

- Assumption that there is an *ideal answer* that contains all relevant documents regarding Q
- Wanted: representation of this answer
- Problem: We do not know it!
- Approximation via relevance feedback to the representation of the ideal answer
 - ▶ representation of the ideal answer contains either 1 or 0 per dimension

Properties of the ideal answer



Further assumptions

- Probability that D is relevant for Q depends only on query representation and document representation
- There exists an ideal answer $R^+(Q)$
- **Chance** that D is relevant with respect to Q :

$$\frac{P(D \in R^+(Q))}{P(D \notin R^+(Q))}$$

Chance

- If probability for relevance becomes 0, then chance is also 0
- For relevance probability of 1, chance goes to infinity
- Chance grows monotonically
 - ▶ i.e. it does not matter whether sorting is done by chance or probability

$$\frac{P(D \in R^+(Q))}{P(D \notin R^+(Q))}$$

8.3 Formal Definition

- Complement to $R^+(Q)$ is $R^-(Q)$: binary relevance decision
 - ▶ i.e. union are all documents
- Weights of representation vectors are binary, i.e. $w_{dk} \in \{0, 1\}$
 - ▶ applies to documents and query

Ranking criterion

- Similarity between D and Q

$$\text{sim}(D, Q) = \frac{P(D' \in R^+(Q) | \beta(D') = \beta(D))}{P(D' \in R^-(Q) | \beta(D') = \beta(D))}$$

- Chance that D' is relevant if it has the same representation as D

Bayes' theorem

- Deals with conditional probabilities, e.g. a under condition b

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

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

Example for Bayes' theorem

- $P(b) = 10\%$
- $P(a \wedge b) = 5\%$
- $P(a|b) = \frac{0.05}{0.1} = 0.5 \Rightarrow 50\%$

Bayes transformations

- Sometimes it can be convenient to infer probabilities from probabilities that are easy to determine

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

$$P(a) \cdot P(b|a) = P(a \wedge b)$$

Bayes transformations

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

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

Insertion into chance

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

a

b

$$\text{sim}(D, Q) = \frac{P(D' \in R^+(Q) | \beta(D') = \beta(D))}{P(D' \in R^-(Q) | \beta(D') = \beta(D))}$$

$$\begin{aligned} \text{sim}(D, Q) &= \frac{P(D' \in R^+(Q)) \cdot P(\beta(D') = \beta(D) | D' \in R^+(Q))}{P(\beta(D') = \beta(D))} \\ &\quad \cdot \frac{P(D' \in R^-(Q)) \cdot P(\beta(D') = \beta(D) | D' \in R^-(Q))}{P(\beta(D') = \beta(D))} \\ &= \frac{P(D' \in R^+(Q)) \cdot P(\beta(D') = \beta(D) | D' \in R^+(Q))}{P(D' \in R^-(Q)) \cdot P(\beta(D') = \beta(D) | D' \in R^-(Q))} \end{aligned}$$

Further simplification

- Probability that any document is (ir)relevant remains constant per query

$$\begin{aligned}
 \text{sim}(D, Q) &= \frac{P(D' \in R^+(Q)) \cdot P(\beta(D') = \beta(D) | D' \in R^+(Q))}{P(\beta(D') = \beta(D))} \\
 &\quad \cdot \frac{P(D' \in R^-(Q)) \cdot P(\beta(D') = \beta(D) | D' \in R^-(Q))}{P(\beta(D') = \beta(D))} \\
 &= \frac{P(D' \in R^+(Q)) \cdot P(\beta(D') = \beta(D) | D' \in R^+(Q))}{P(D' \in R^-(Q)) \cdot P(\beta(D') = \beta(D) | D' \in R^-(Q))}
 \end{aligned}$$

Preliminary final result

$$\text{sim}(D, Q) \approx \frac{P(\beta(D') = \beta(D) | D' \in R^+(Q))}{P(\beta(D') = \beta(D) | D' \in R^-(Q))}$$

- Numerator:
 - ▶ probability that representation is given if the document is relevant
- Denominator:
 - ▶ probability that representation is given if it is irrelevant

Example

	D1	D2	D3	D4	D5	D6	D7	D8	D9
w_{d1}	1	1	1	0	0	0	1	1	0
w_{d2}	1	0	1	0	1	0	1	0	0
Relevance	✓	✓	✓	✗	✗	✗	✗	✗	✗

- New document $D_{11} = (1, 1)$, $\text{sim}(D_{11}, Q) = 4$

$$\text{sim}(D, Q) \approx \frac{P(\beta(D') = \beta(D) | D' \in R^+(Q))}{P(\beta(D') = \beta(D) | D' \in R^-(Q))}$$

2
3
1
6

Example

	D1	D2	D3	D4	D5	D6	D7	D8	D9
w_{d1}	1	1	1	0	0	0			
w_{d2}	1	0	1	0	1				
Relevance	✓	✓	✓	✗					✗

- New document

**Still the open problem:
How to determine enough samples with 2^t
possibilities?**

$$P(\beta(D') = \beta(D) | D' \in R^+(Q)) \approx \frac{P(\beta(D') = \beta(D) | D' \in R^+(Q))}{P(\beta(D') = \beta(D) | D' \in R^-(Q))}$$

2
3
1
6

Componentwise evaluation

- Probabilities are no longer determined per description vector, but per vector component
- We replace $\beta(D') = \beta(D)$ with componentwise equality

$$\text{sim}(D, Q) \approx \frac{P(\forall_{\{i|w_{di}=1\}}(w_{d'i}=1) \wedge \forall_{\{i|w_{di}=0\}}(w_{d'i}=0) | D' \in R^+(Q))}{P(\forall_{\{i|w_{di}=1\}}(w_{d'i}=1) \wedge \forall_{\{i|w_{di}=0\}}(w_{d'i}=0) | D' \in R^-(Q))}$$

The “Independence” in the BIR model

$$\text{sim}(D, Q) \approx \frac{P(\forall_{\{i|w_{di}=1\}}(w_{d'i}=1) \wedge \forall_{\{i|w_{di}=0\}}(w_{d'i}=0) | D' \in R^+(Q))}{P(\forall_{\{i|w_{di}=1\}}(w_{d'i}=1) \wedge \forall_{\{i|w_{di}=0\}}(w_{d'i}=0) | D' \in R^-(Q))}$$

- Assuming that all components are independent of each other, then:
 $P(a \wedge b) = P(a) \cdot P(b)$

$$\begin{aligned} \text{sim}(D, Q) &\approx \frac{(\prod_{\{i|w_{di}=1\}} P(w_{d'i}=1) | D' \in R^+(Q))}{(\prod_{\{i|w_{di}=1\}} P(w_{d'i}=1) | D' \in R^-(Q))} \\ &\cdot \frac{(\prod_{\{i|w_{di}=0\}} P(w_{d'i}=0) | D' \in R^+(Q))}{(\prod_{\{i|w_{di}=0\}} P(w_{d'i}=0) | D' \in R^-(Q))} \end{aligned}$$

Further transformations

- Goal: avoid multiplying all terms
- We know:
 - ▶ $\log(a_1 \cdot a_2 \cdot a_3) = \log(a_1) + \log(a_2) + \log(a_3)$
 - ▶ if $a_1 > a_2 > a_3$, then this is also true for their logarithms

Final formula

- Under the following assumptions:

- ▶ binary relevance judgment ($P(R^+) + P(R^-) = 1$)
- ▶ we consider only query terms = 1
- ▶ we can omit all factors that are the same or assumed to be the same for all documents related to the query
- ▶ we obtain:

$$\begin{aligned} \text{sim}(D, Q) \approx \sum_{\{i | w_{qi} = w_{di} = 1\}} & \left(\log \frac{P(w_{d'i} = 1) | D' \in R^+(Q)}{1 - P(w_{d'i} = 1) | D' \in R^+(Q)} \right. \\ & \left. + \log \frac{1 - P(w_{d'i} = 1) | D' \in R^-(Q)}{P(w_{d'i} = 1) | D' \in R^-(Q)} \right) \end{aligned}$$

Determination of the initial result

- Searched: estimate that $P(w_{d'i} = 1 | D' \in R^+(Q))$ or $P(w_{d'i} = 1 | D' \in R^-(Q))$
- Assumption: equally frequent, i.e.:
 - ▶ $P(w_{d'i} = 1 | D' \in R^+(Q)) = 0.5$
 - ▶ $P(w_{d'i} = 1 | D' \in R^-(Q)) = \frac{n_i}{N}$

Determination of the initial result

$$\begin{aligned} \text{sim}(D, Q) &\approx \sum_{\{i | w_{qi}=w_{di}=1\}} \left(\log \frac{0.5}{0.5} + \log \frac{1 - \frac{n_i}{N}}{\frac{n_i}{N}} \right) \\ &= \sum_{\{i | w_{qi}=w_{di}=1\}} \log \frac{N - n_i}{n_i} \end{aligned}$$

- Strong similarity to the $tf \cdot idf$ formula
 - ▶ idf : $N - n_i \approx N$ for large N
 - ▶ tf : because of sum index and binary values

Example

- Index vocabulary $t = 8$
- Query vector $Q = (1, 0, 0, 1, 1, 0, 0, 0)$
- Number of documents $N = 500$
- $n_1 = 87$; $n_4 = 23$; $n_5 = 100$
- Document $D = (1, 0, 0, 0, 1, 1, 0, 0)$

$$\text{sim}(D, Q) = \log\left(\frac{500 - 87}{87}\right) + \log\left(\frac{500 - 100}{100}\right) \approx 1.28$$

8.4 Relevance Feedback

- Idea of relevance feedback
 - ▶ pseudo relevance
 - ★ feedback does not require user interaction
 - ★ let V be the set of r highest ranked documents
 - ▶ real relevance feedback
 - ★ let V^* be the set of positively evaluated documents

Adjustment according to relevance feedback

$$P(w_{d'i} = 1 | D' \in R^+(Q)) = \frac{|\{D' \in V | w_{d'i} = 1\}|}{|V|}$$

$$P(w_{d'i} = 1 | D' \in R^-(Q)) = \frac{n_i - |\{D' \in V | w_{d'i} = 1\}|}{N - |V|}$$

Variants for very small sets V

$$P(w_{d'i} = 1 | D' \in R^+(Q)) = \frac{|\{D' \in V | w_{d'i} = 1\}| + 0.5}{|V| + 1}$$

$$P(w_{d'i} = 1 | D' \in R^-(Q)) = \frac{n_i - |\{D' \in V | w_{d'i} = 1\}| + 0.5}{N - |V| + 1}$$

Or

$$P(w_{d'i} = 1 | D' \in R^+(Q)) = \frac{|\{D' \in V | w_{d'i} = 1\}| + \frac{n_i}{N}}{|V| + 1}$$

$$P(w_{d'i} = 1 | D' \in R^-(Q)) = \frac{n_i - |\{D' \in V | w_{d'i} = 1\}| + \frac{n_i}{N}}{N - |V| + 1}$$

Criticism of the BIR model

- Independence assumption of the individual components (as with the vector space model)
- Initial result is estimated very roughly, relevance feedback becomes necessary
 - ▶ improvement only valid for one query, no general learning
- Binary weighting is less informative than weighting in vector space model

Experiments

method used		CACM 1033 Doc. 30 queries	CISI 12684 Doc. 84 queries	CRAN 1397 Doc. 225 queries	INSPEC 1460 Doc. 112 queries	MED 3204 Doc. 64 queries	average
initial query							
	Precision	0,1459	0,1184	0,1156	0,1368	0,3346	
IDE (dec hi)							
with all terms	Precision	0,2704	0,1742	0,3011	0,2140	0,6305	
	improvement	+86%	+47%	+160%	+56%	+88%	+87%
selected terms	Precision	0,2479	0,1924	0,2498	0,1976	0,6218	
	improvement	+70%	+63%	+116%	+44%	+86%	+76%
BIR model							
with all terms	Precision	0,2289	0,1436	0,3108	0,1621	0,5972	
	improvement	+57%	+21%	+169%	+19%	+78%	+69%
selected terms	Precision	0,2224	0,1634	0,2120	0,1876	0,5643	
	improvement	+52%	+38%	+83%	+37%	+69%	+56%

8.5 BM25

- Developed by Robertson/Sparck Jones
- First used in the Okapi system at City University (London)
- Successfully used on TREC
- Based on probabilistic retrieval
- Combines $tf \cdot idf$ variant with probabilistic retrieval


BM25

$$\text{sim}(D, Q) = \sum_{i=1}^t \text{IDF}(n_i) \cdot \frac{(k_1 + 1)tf_{di}}{k_1((1 - b) + b\frac{dl}{avdl}) + tf_{di}} \cdot \frac{(k_3 + 1)tf_{qi}}{k_3 + tf_{qi}}$$
$$\text{IDF}(n_i) = \log \frac{N - n_i + 0.5}{n_i + 0.5}$$


- k_1 (between 1.0 and 2.0), b (normally 0.75) and k_3 (between 0 and 1000) are constants
- dl : length of the document in words
- $avdl$: average length of documents

BM25 vs. vector space model

$$sim(D, Q) = \sum_{i=1}^t IDF(n_i) \cdot \frac{(k_1 + 1)tf_{di}}{k_1((1 - b) + b\frac{dl}{\overline{avdl}}) + tf_{di}} \cdot \frac{(k_3 + 1)tf_{qi}}{k_3 + tf_{qi}}$$

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


- Attention! If $n_i > \frac{N}{2}$, IDF can become negative:
 - ▶ e.g. set to zero



$$w_{dk} = \frac{tf_{dk} \log \frac{N}{n_k}}{\sqrt{\sum_{i=1}^t (tf_{di} \log \frac{N}{n_i})^2}}$$

Slope

- In the vector space model, the disadvantage of short documents should be eliminated by normalization
- Problem: short documents contain few terms, i.e. their description vectors have few components with $w_{dk} > 0$

Slope

- Result:

- ▶ short documents have large component values after normalization, since normalized length 1 is “divided” among few components
- ▶ long documents have small component values
- ▶ \Rightarrow with few query terms, documents with high w_{dk} values are preferred (short documents)

Slope

- Old normalization $\sqrt{\sum_{i=1}^t (tf_{di} \cdot \log \frac{N}{n_i})^2} = \sqrt{\sum_{i=1}^t w_{di}^2}$ is replaced by:

$$kf_d = (1 - slope) + slope \cdot \frac{old_normalization}{average_old_normalization}$$

- Slope mostly in the range 0.65 – 0.75 (determine empirically)
- New $w_{dk} = kf_d \cdot w_{dk}$

Slope at BM25

$$sim(D, Q) = \sum_{i=1}^t IDF(n_i) \cdot \frac{(k_1 + 1)tf_{di}}{k_1((1 - b) + b\frac{dl}{avdl}) + tf_{di}} \cdot \frac{(k_3 + 1)tf_{qi}}{k_3 + tf_{qi}}$$

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

- b and k_1 control the slope cf. vector space model

$$kf_d = (1 - slope) + slope \cdot \frac{old_normalization}{average_old_normalization}$$