Chapter IR:III

III. Retrieval Models

- Overview of Retrieval Models
- Boolean Retrieval
- Vector Space Model
- Binary Independence Model
- Okapi BM25
- Divergence From Randomness
- Latent Semantic Indexing
- Explicit Semantic Analysis
- Language Models

- Combining Evidence
- Learning to Rank

Overview of Retrieval Models

Document Views

Information retrieval requires modeling and representing documents on a computer. We distinguish three orthogonal views on a document's content:

1. Layout view
   Presentation of a document on a (two-dimensional) medium.

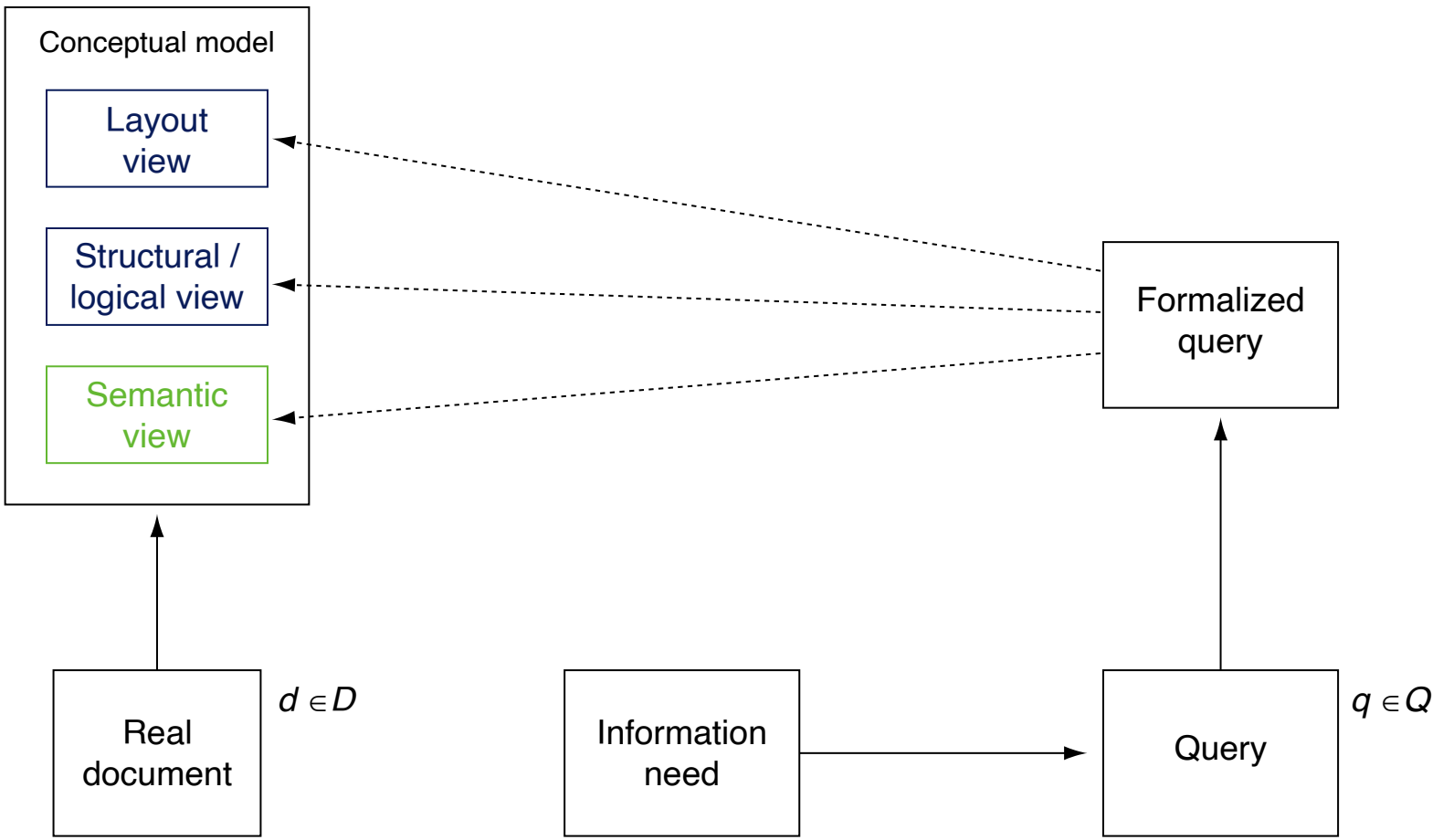
2. Structural / logical view
   Composition and logical structure of a document. Example:

```
\documentclass[twocolumn,english]{article}
\title{...}
\author{...}
\section{...}
```

3. Semantic view
   The meaning of a document or its message, allowing for its interpretation.

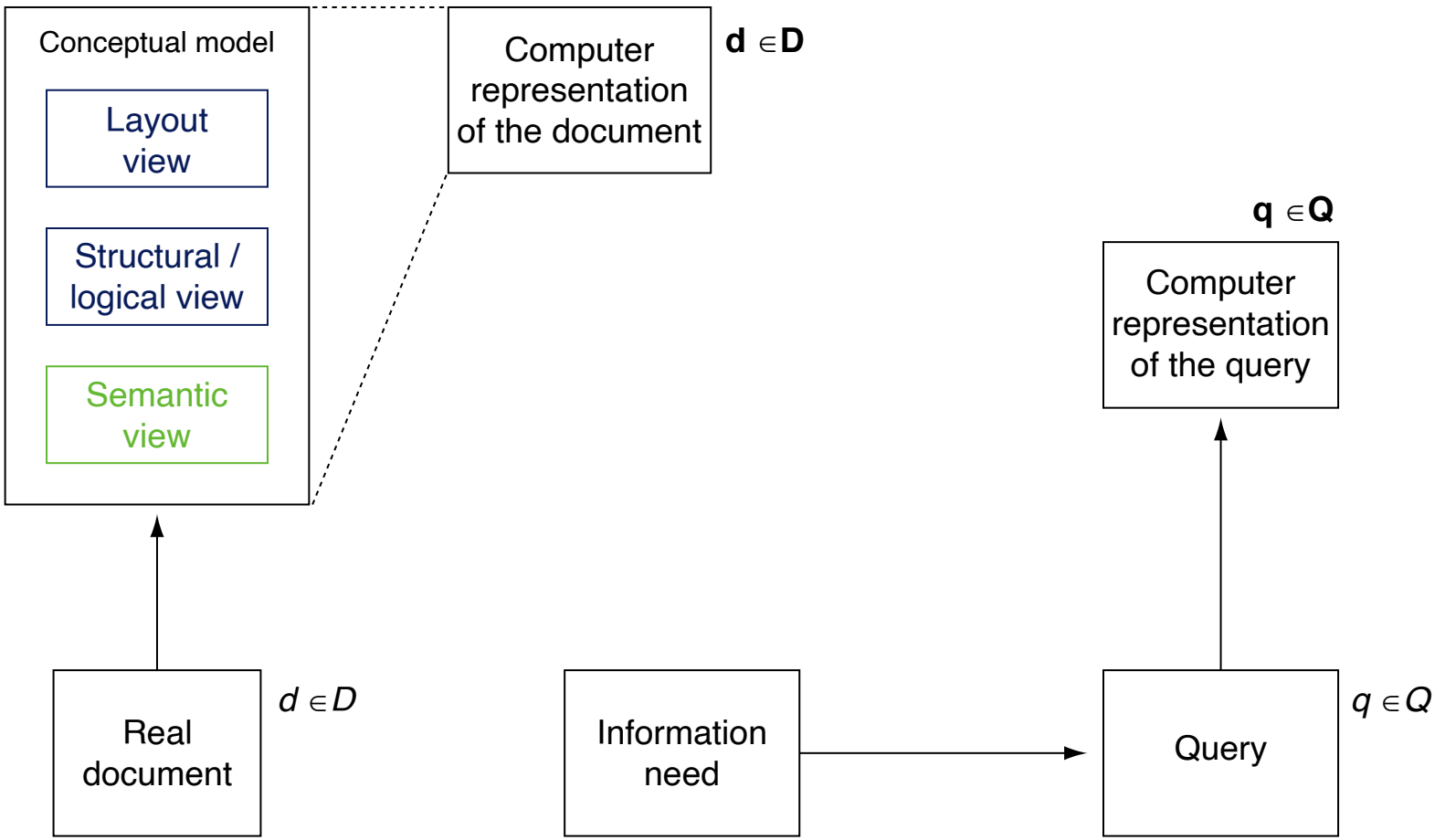
Overview of Retrieval Models

Retrieval Models



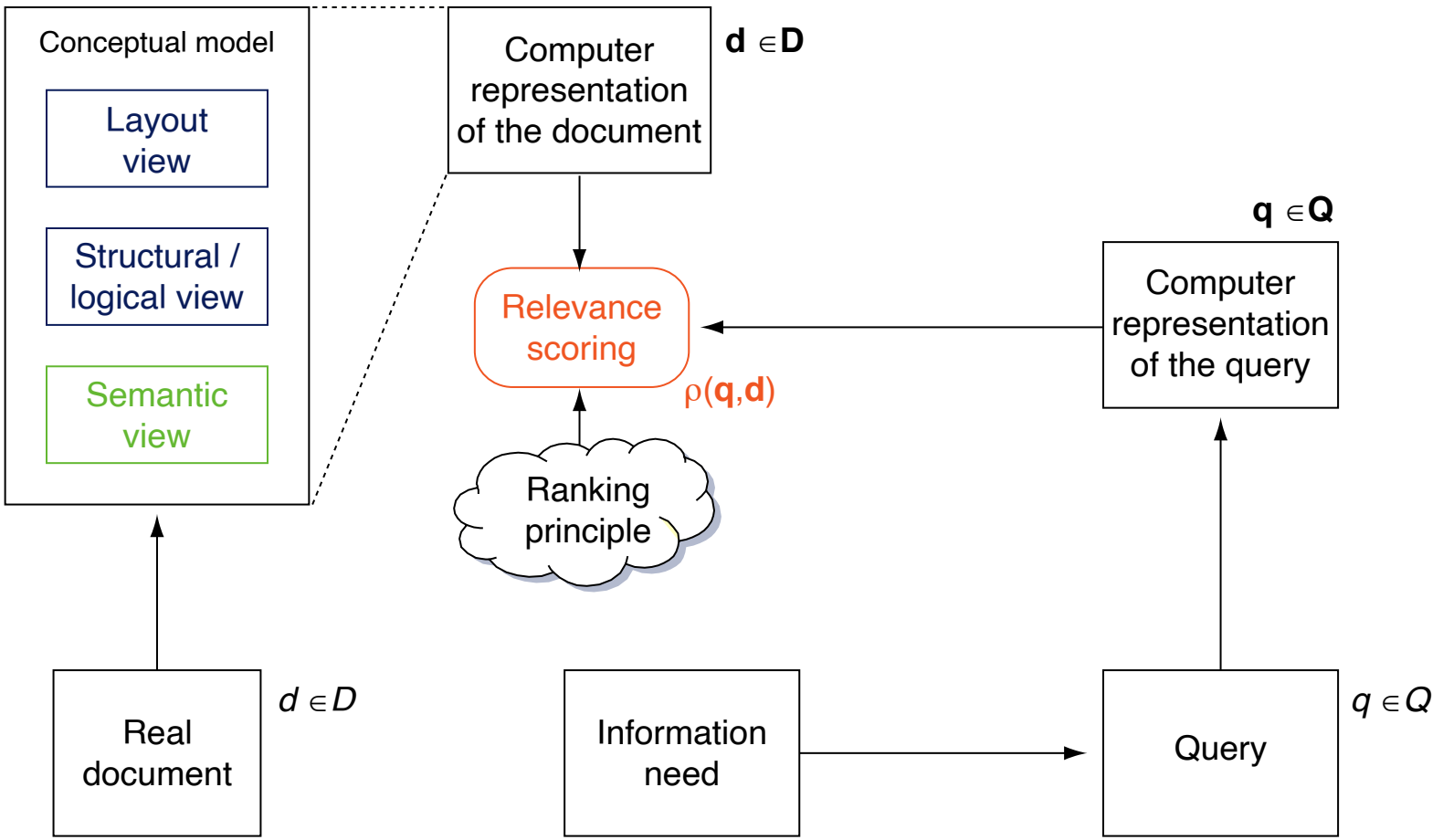
Overview of Retrieval Models

Retrieval Models



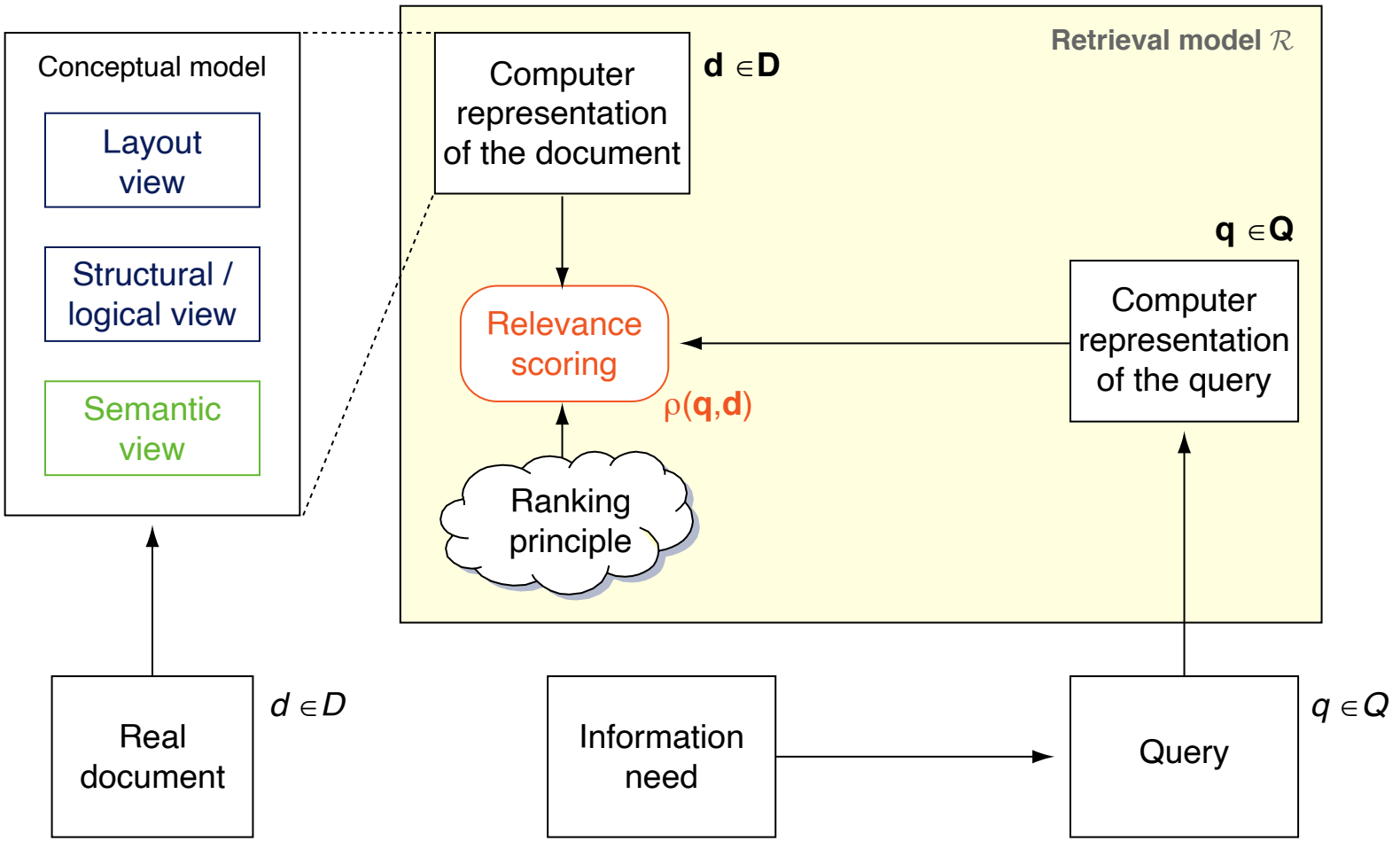
Overview of Retrieval Models

Retrieval Models



Overview of Retrieval Models

Retrieval Models



Overview of Retrieval Models

Definition 1 (Retrieval Model, Relevance Function)

Let D denote the set of documents and Q the set of queries. A retrieval model $\mathcal{R}$ for D, Q is a tuple $\langle \mathbf{D}, \mathbf{Q}, \rho \rangle$ defined as follows:

1. $\mathbf{D}$ is the set of document representations, where $\mathbf{d} \in \mathbf{D}$ represents $d \in D$.
   It may encode information from the layout view, the logical view, and the semantic view.

2. $\mathbf{Q}$ is the set of query representations.

Overview of Retrieval Models

Definition 1 (Retrieval Model, Relevance Function)

Let D denote the set of documents and Q the set of queries. A retrieval model $\mathcal{R}$ for D, Q is a tuple $\langle \mathbf{D}, \mathbf{Q}, \rho \rangle$ defined as follows:

1. $\mathbf{D}$ is the set of document representations, where $\mathbf{d} \in \mathbf{D}$ represents $d \in D$.
   It may encode information from the layout view, the logical view, and the semantic view.

2. $\mathbf{Q}$ is the set of query representations.

3. $\rho(\mathbf{q}, \mathbf{d})$ denotes a relevance function, which quantifies the relevance between a query q and a document d via their representations $\mathbf{q} \in \mathbf{Q}$ and $\mathbf{d} \in \mathbf{D}$:

   $$\rho : \mathbf{Q} \times \mathbf{D} \to \mathbf{R}$$

   The values computed by ρ are called relevance scores.

$\mathcal{R}$ formalizes a certain ranking principle.

Remarks:

- A document representation encompasses certain elements and specific aspects of a real document. Examples for document representations include feature vectors and fingerprints.
- A retrieval model provides the theoretical foundations of how human information needs can be satisfied by drawing information from the three views. Examples for retrieval models include the vector space model, the binary independence model, and latent semantic indexing.
- An alternative name for a retrieval model is retrieval strategy.
- Most retrieval models are based on the semantic view of documents.
- An intensional definition of the sets $\mathbf{Q}$ and $\mathbf{D}$ can be given as functions $\alpha_Q : Q \to \mathbf{Q}$ and $\alpha_D : D \to \mathbf{D}$. [Fuhr 2004]

Overview of Retrieval Models

Probability Ranking Principle (PRP) [Robertson 1977, 2009]

If retrieved documents are ordered by decreasing probability of relevance on the data available, then the system's effectiveness is the best that can be obtained for the data.

Overview of Retrieval Models

Probability Ranking Principle (PRP) [Robertson 1977, 2009]

> If retrieved documents are ordered by decreasing probability of relevance on the data available, then the system's effectiveness is the best that can be obtained for the data.

Assumptions:

- Relevance *rel*(d, q) is a property of a document d given an information need's query q, assessable without reference to other documents.
- Relevance is binary: *rel*$(d, q) \in \{0, 1\}$.
- Relevance is that of an individual user submitting q.

Ranking by probability of relevance provably maximizes several objective functions:

- Expected recall
- Expected precision
- Expected utility

Remarks:

- The probability ranking principle has not been shown to hold in general as of yet, but under the above assumptions.
- In a counterexample, William S. Cooper considers different users with different information needs who formulate the same query.
- Including more knowledge about the user accounts takes user relevance into account: $P(\textit{rel}(d, q) = 1 \mid \mathbf{d}, \mathbf{q}, \mathbf{u})$, where $\mathbf{u}$ is a user model.

Overview of Retrieval Models

Ranking Principles in IR

A ranking principle states a criterion and shows that ranking documents by this criterion achieves an objective, usually the maximization of an objective function.

- Binary user relevance $\mathit{rel}(d, q) \in \{0, 1\}$ of a document d to a query q.
  Objective: Return relevant documents.
- Semantic similarity $\varphi(\mathbf{d}, \mathbf{q})$ of a document d to a query q.
  Objective: Return documents that are similar to the topic of the query.
- Probability of user relevance $P(\mathit{rel}(d, q) = 1 \mid \mathbf{d}, \mathbf{q})$ of d to q.
  Objective: Return documents that satisfy the user's information need.
- Probability $P(\mathbf{d} \mid \mathbf{q})$ of d having been generated for the topic of query q.
  Objective: Return documents that fit the topic of the query.
- Amount of information $-\log_2 P(\mathbf{d} \mid \mathit{tf}(t_1), \ldots, \mathit{tf}(t_{|\mathbf{q}|}))$ carried by $\mathbf{q}$'s terms in $\mathbf{d}$.
  Objective: Return documents that carry much information about $\mathbf{q}$'s terms.

Overview of Retrieval Models

Types of Retrieval Models [Amati 2018]

Logical models: Evaluation for a given document d and a query q of the truth of

$$\mathcal{I}(\mathbf{d} \rightarrow \mathbf{q})$$

Algebraic models: Computation of the similarity between d and q in a vector space:

$$\varphi(\mathbf{d}, \mathbf{q})$$

Probabilistic models: Estimation of the probability of a user's relevance *rel* of d for q:

$$P(\mathit{rel}(d, q) = 1 \mid \mathbf{d}, \mathbf{q})$$

Bayesian models: Estimation of the probability of generating d from q:

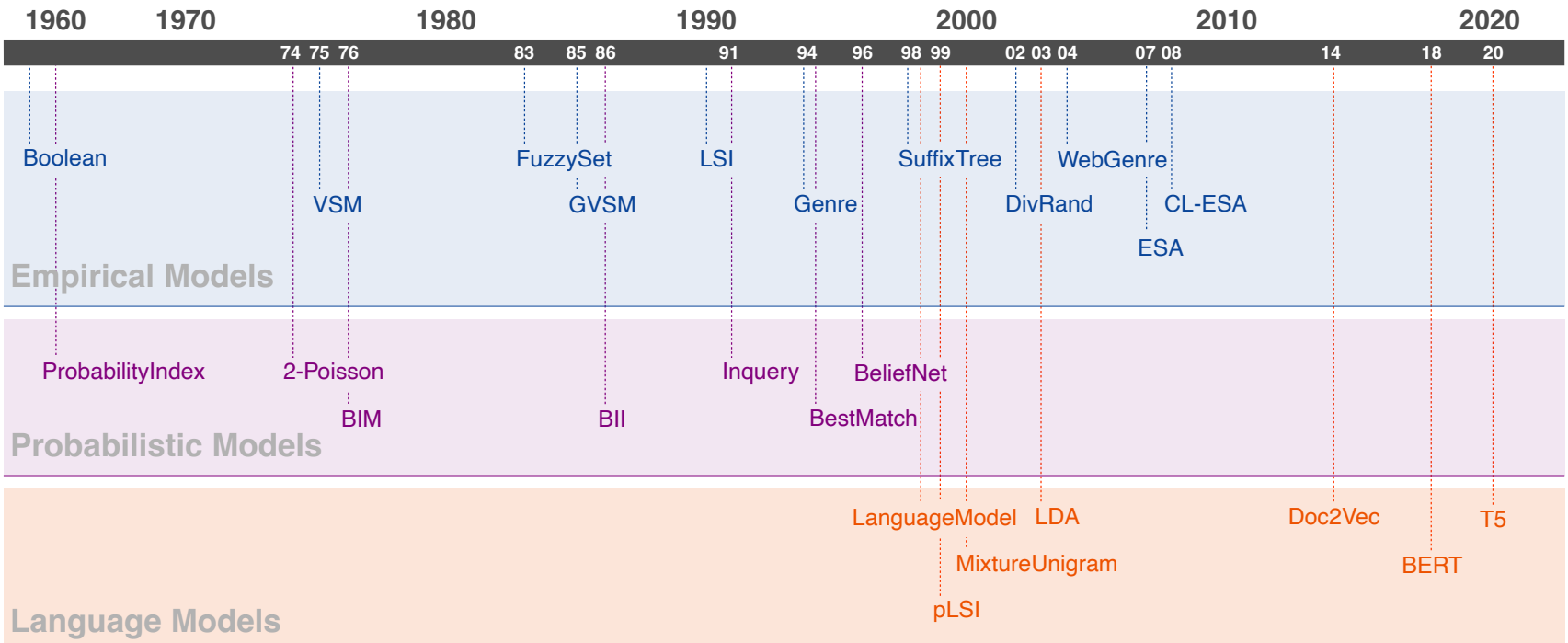
$$P(\mathbf{d} \mid \mathbf{q})$$

Information theoretic models: Computation of the number of bits necessary to code $\mathit{tf}(t_i)$ many of $\mathbf{q}$'s i-th term t_i in $\mathbf{d}$ for $i \in \{1, \ldots, |\mathbf{q}|\}$:

$$-\log_2 P(\mathbf{d} \mid \mathit{tf}(t_1), \ldots, \mathit{tf}(t_{|\mathbf{q}|}))$$

Overview of Retrieval Models

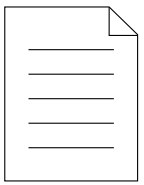
History of Retrieval Models [Stein 2013]



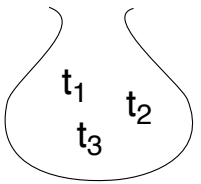
Overview of Retrieval Models

Document Modeling

Analytic models



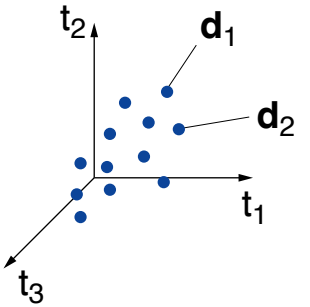
d



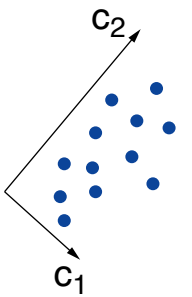
"bag of words" model

	d_1	d_2	⋯
t_1	5	6	
t_2	7	5	⋮
t_3	1	2	

term-document matrix



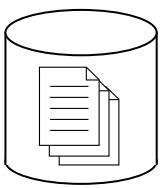
vector space

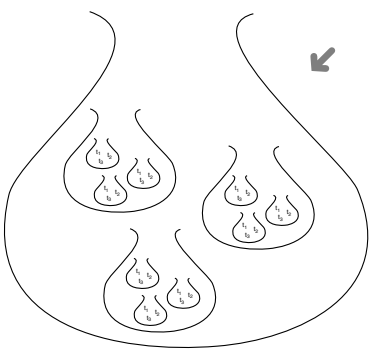


concept space

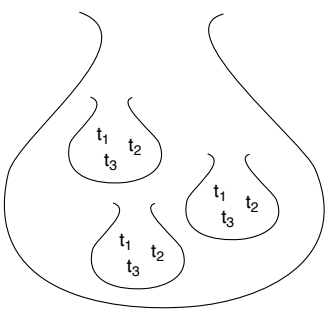
embedding

Synthetic models / Generative models

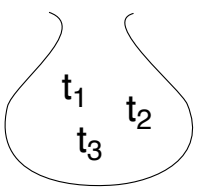




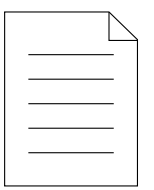
distribution of topic distributions



topic distribution



urn model / topic model



d

Chapter IR:III

III. Retrieval Models

- Overview of Retrieval Models
- Boolean Retrieval
- Vector Space Model
- Binary Independence Model
- Okapi BM25
- Divergence From Randomness
- Latent Semantic Indexing
- Explicit Semantic Analysis
- Language Models
- Combining Evidence
- Learning to Rank

Boolean Retrieval

Retrieval Model $\mathcal{R} = \langle \mathbf{D}, \mathbf{Q}, \rho \rangle$ [Generic Model] [Boolean] [VSM] [BIM] [BM25] [LSI] [ESA] [LM]

Document representations $\mathbf{D}$.

- $T = \{t_1, \ldots, t_m\}$ is the set of m index terms (lemmatized or stemmed words).
- T are the atoms of a logical formula for d with operators $\wedge$, $\vee$, $\neg$, and brackets.
- $\mathbf{d} = (\bigwedge_{t \in d} t) \wedge \neg(\bigwedge_{t \notin d} t)$, where $\mathcal{I}_{\mathbf{d}}(t) = 1$ if t occurs in d, and $\mathcal{I}_{\mathbf{d}}(t) = 0$ otherwise.

Query representations $\mathbf{Q}$.

- $\mathbf{q}$ is a logical formula over T.

Relevance function ρ.

- $\rho(d, q) = \mathcal{I}(\mathbf{d} \rightarrow \mathbf{q})$, where $\rightarrow$ is the logical implication.
- $\rho(d, q) = 1$ indicates relevance of d to q, and $\rho(d, q) = 0$ otherwise.
- $R_q \subseteq D$ is the set of documents $d \in D$ relevant to q, i.e., with $\rho(d, q) = 1$.
- $\rho'(d, q) = P(\mathcal{I}(\mathbf{d} \rightarrow \mathbf{q}) = 1) = P(\mathbf{d} \rightarrow \mathbf{q}) = P(q \mid d)$ relaxes relevance scoring.

Boolean Retrieval

Retrieval Model $\mathcal{R} = \langle \mathbf{D}, \mathbf{Q}, \rho \rangle$ [Generic Model] [Boolean] [VSM] [BIM] [BM25] [LSI] [ESA] [LM]

Document representations $\mathbf{D}$.

- $T = \{t_1, \ldots, t_m\}$ is the set of m index terms (lemmatized or stemmed words).
- T are the atoms of a logical formula for d with operators $\wedge$, $\vee$, $\neg$, and brackets.
- $\mathbf{d} = (\bigwedge_{t \in d} t) \wedge \neg(\bigwedge_{t \notin d} t)$, where $\mathcal{I}_{\mathbf{d}}(t) = 1$ if t occurs in d, and $\mathcal{I}_{\mathbf{d}}(t) = 0$ otherwise.

Query representations $\mathbf{Q}$.

- $\mathbf{q}$ is a logical formula over T.

Relevance function ρ.

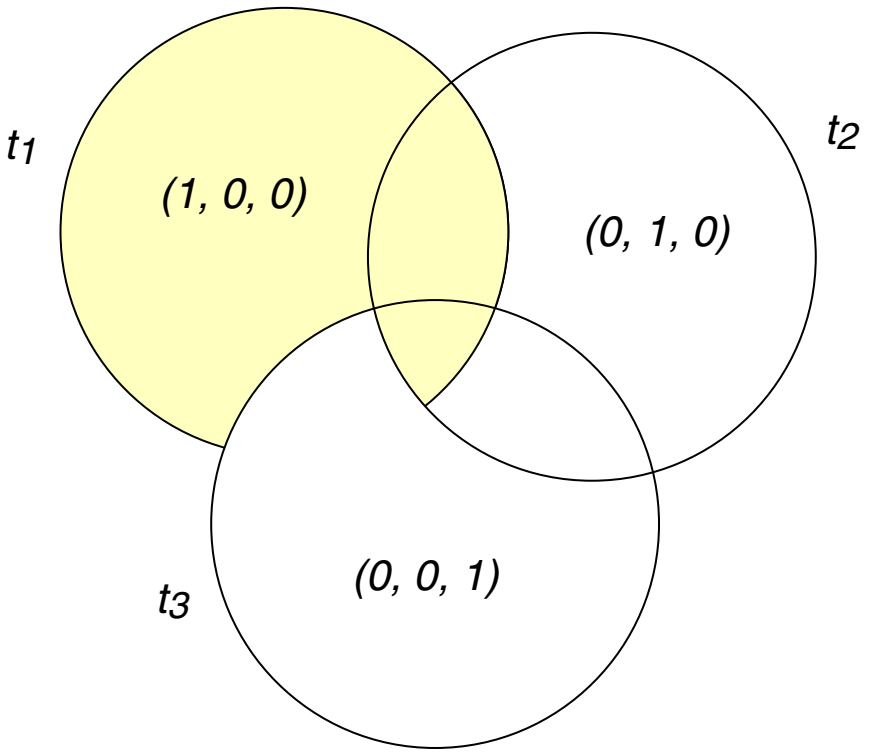
- $\rho(d, q) = \mathcal{I}(\mathbf{d} \rightarrow \mathbf{q})$, where $\rightarrow$ is the logical implication.
- $\rho(d, q) = 1$ indicates relevance of d to q, and $\rho(d, q) = 0$ otherwise.
- $R_q \subseteq D$ is the set of documents $d \in D$ relevant to q, i.e., with $\rho(d, q) = 1$.
- $\rho'(d, q) = P(\mathcal{I}(\mathbf{d} \rightarrow \mathbf{q}) = 1) = P(\mathbf{d} \rightarrow \mathbf{q}) = P(q \mid d)$ relaxes relevance scoring.

Remarks:

- $\mathcal{I} : T \to \{0, 1\}$ and $\mathcal{I} : \{\alpha \mid \alpha \text{ is a logical formula over } T\} \to \{0, 1\}$ is the evaluation or interpretation function that assigns truth values to the atoms T as well as to propositional formulas over them.

Boolean Retrieval

Relevance Function ρ



What query is illustrated?

Boolean Retrieval

Example

Document representation:

$$\mathbf{d} = \texttt{chrysler} \wedge \texttt{deal} \wedge \texttt{usa} \wedge \texttt{china} \wedge \neg\texttt{cat} \wedge \texttt{sales} \wedge \neg\texttt{dog} \wedge \ldots$$

Query representation:

$$\begin{aligned}\mathbf{q} &= \texttt{usa} \wedge (\texttt{dog} \vee \neg\texttt{cat}) \\ &\equiv (\texttt{usa} \wedge \texttt{dog}) \vee (\texttt{usa} \wedge \neg\texttt{cat}) \\ &\equiv (\texttt{usa} \wedge \neg\texttt{dog} \wedge \neg\texttt{cat}) \vee (\texttt{usa} \wedge \texttt{dog} \wedge \neg\texttt{cat}) \vee (\texttt{usa} \wedge \texttt{dog} \wedge \texttt{cat})\end{aligned}$$

Relevance function:

$\rho(d, q) = \mathcal{I}(\mathbf{d} \to \mathbf{q}) = 1$, since $\mathcal{I}_{\mathbf{d}}(\texttt{usa}) = 1, \mathcal{I}_{\mathbf{d}}(\texttt{dog}) = 0$, and $\mathcal{I}_{\mathbf{d}}(\texttt{cat}) = 0$.

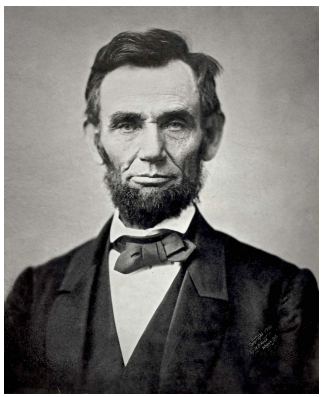
Remarks:

- The symbol "$\equiv$" denotes "is logically equivalent with".
- What does logical equivalence mean?
- A Boolean query in disjunctive normal form can be answered straightforward using an inverted index in parallel for each conjunction.
- A Boolean query in canonical disjunctive normal form will retrieve each document only once.

Boolean Retrieval

Query Refinement: “Searching by Numbers”

Best practice in Boolean retrieval: (re)formulate queries until the number of documents retrieved is manageable. Example: pages about President Lincoln.



1. `lincoln`
   Results: many pages about cars, places, people
2. `president ∧ lincoln`
   A result: “Ford Motor Company today announced that Darryl Hazel will succeed Brian Kelley as president of Lincoln Mercury.”
3. `president ∧ lincoln ∧ ¬automobile ∧ ¬car`
   Not a result: “President Lincoln’s body departs Washington in a nine-car funeral train.”
4. `president ∧ lincoln ∧ ¬automobile ∧ biography ∧ life ∧ birthplace ∧ gettysburg`
   Results: ∅
5. `president ∧ lincoln ∧ ¬automobile ∧ (biography ∨ life ∨ birthplace ∨ gettysburg)`
   A result: “President’s Day – Holiday activities – crafts, mazes, word searches, . . . ’The Life of Washington’ Read the entire book online! Abraham Lincoln Research Site”

Boolean Retrieval

Discussion

Advantages:

- Precision: in principle, any subset of documents from a collection can be designated by a Boolean query
- as in data retrieval, other fields are possible (e.g., date, document type, etc.)
- simple, efficient implementation

Disadvantages:

- retrieval effectiveness depends entirely on the user
- cumbersome query formulation (e.g., expertise required)
- no possibility to weight query terms
- no ranking; binary relevance scoring is too restrictive for most practical purposes (exceptions: medical retrieval, patent retrieval, eDiscovery (law))
- the size of the result set is difficult to be controlled

Vector Space Model

Retrieval Model $\mathcal{R} = \langle \mathbf{D}, \mathbf{Q}, \rho \rangle$ [Generic Model] [Boolean] [VSM] [BIM] [BM25] [LSI] [ESA] [LM]

Document representations $\mathbf{D}$.

- $T = \{t_1, \ldots, t_m\}$ is the set of m index terms (word stems, without stop words).
- T is interpreted as set of dimensions of an m-dimensional vector space.
- $\omega : \mathbf{D} \times T \to \mathbf{R}$ is a term weighting function, quantifying term importance.
- $\mathbf{d} = (w_1, \ldots, w_m)^T$, where $w_i = \omega(\mathbf{d}, t_i)$ is the term weight of the i-th term in T.

Query representations $\mathbf{Q}$.

- $\mathbf{q} = (w_1, \ldots, w_m)^T$, where $w_i = \omega(\mathbf{q}, t_i)$ is the term weight of the i-th term in T.

Relevance function ρ.

- Distance and similarity functions φ serve as relevance functions.
- $\rho(d, q) = \varphi(\mathbf{d}, \mathbf{q}) = \mathbf{d}^T\mathbf{q}$, the scalar product of vectors $\mathbf{d}$ and $\mathbf{q}$.
- Normalizing $\mathbf{d}$ and $\mathbf{q}$ calculates cosine similarity.

Vector Space Model

Retrieval Model $\mathcal{R} = \langle \mathbf{D}, \mathbf{Q}, \rho \rangle$ [Generic Model] [Boolean] [VSM] [BIM] [BM25] [LSI] [ESA] [LM]

Document representations $\mathbf{D}$.

- $T = \{t_1, \ldots, t_m\}$ is the set of m index terms (word stems, without stop words).
- T is interpreted as set of dimensions of an m-dimensional vector space.
- $\omega : \mathbf{D} \times T \to \mathbf{R}$ is a term weighting function, quantifying term importance.
- $\mathbf{d} = (w_1, \ldots, w_m)^T$, where $w_i = \omega(\mathbf{d}, t_i)$ is the term weight of the i-th term in T.

Query representations $\mathbf{Q}$.

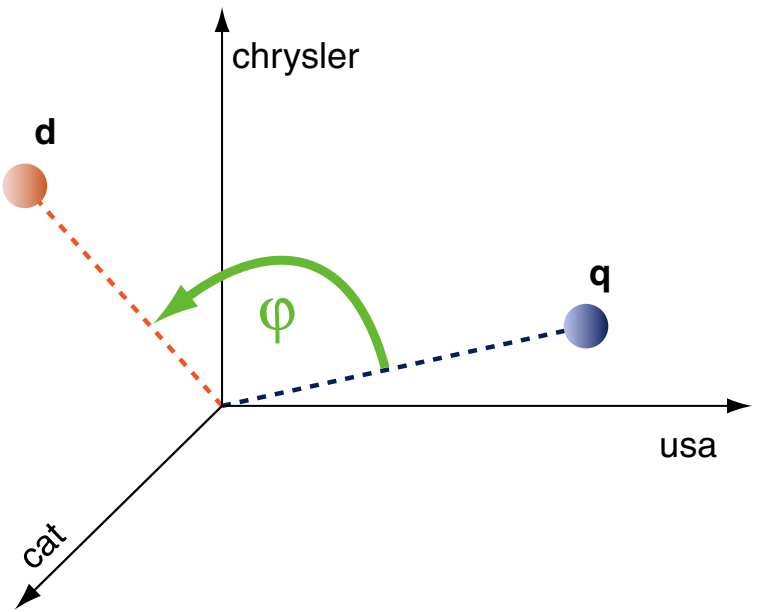
- $\mathbf{q} = (w_1, \ldots, w_m)^T$, where $w_i = \omega(\mathbf{q}, t_i)$ is the term weight of the i-th term in T.

Relevance function ρ.

- Distance and similarity functions φ serve as relevance functions.
- $\rho(d, q) = \varphi(\mathbf{d}, \mathbf{q}) = \mathbf{d}^T\mathbf{q}$, the scalar product of vectors $\mathbf{d}$ and $\mathbf{q}$.
- Normalizing $\mathbf{d}$ and $\mathbf{q}$ calculates cosine similarity.

Vector Space Model

Relevance Function ρ: Cosine Similarity



Vector Space Model

Relevance Function ρ: Cosine Similarity

The scalar product $\mathbf{a}^T\mathbf{b}$ between two m-dimensional vectors $\mathbf{a}$ and $\mathbf{b}$, where φ denotes the angle between them, is defined as follows:

$$\begin{aligned} \mathbf{a}^T\mathbf{b} &= ||\mathbf{a}|| \cdot ||\mathbf{b}|| \cdot \cos(\varphi) \\ \Leftrightarrow \cos(\varphi) &= \frac{\mathbf{a}^T\mathbf{b}}{||\mathbf{a}|| \cdot ||\mathbf{b}||}, \end{aligned}$$

where $||\mathbf{x}||$ denotes the L2 norm of vector $\mathbf{x}$:

$$||\mathbf{x}|| = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}$$

Let $\rho(\mathbf{q}, \mathbf{d}) = \cos(\varphi)$ be the relevance function of the vector space model.

Vector Space Model

Example

$$\mathbf{d} = \begin{pmatrix} \texttt{chrysler} & w_1 \\ \texttt{usa} & w_2 \\ \texttt{cat} & w_3 \\ \texttt{dog} & w_4 \\ \texttt{mouse} & w_5 \end{pmatrix} = \begin{pmatrix} \texttt{chrysler} & 1 \\ \texttt{usa} & 4 \\ \texttt{cat} & 3 \\ \texttt{dog} & 7 \\ \texttt{mouse} & 5 \end{pmatrix}$$

$$\mathbf{d}' = \begin{pmatrix} \texttt{chrysler} & 0.05 \\ \texttt{usa} & 0.2 \\ \texttt{cat} & 0.15 \\ \texttt{dog} & 0.35 \\ \texttt{mouse} & 0.25 \end{pmatrix}, \qquad \mathbf{q}' = \begin{pmatrix} \texttt{chrysler} & 0.2 \\ \texttt{usa} & 0.2 \\ \texttt{cat} & 0.2 \\ \texttt{dog} & 0.2 \\ \texttt{elephant} & 0.2 \end{pmatrix}$$

Vector Space Model

Example

$$\mathbf{d} = \begin{pmatrix} \texttt{chrysler} & w_1 \\ \texttt{usa} & w_2 \\ \texttt{cat} & w_3 \\ \texttt{dog} & w_4 \\ \texttt{mouse} & w_5 \end{pmatrix} = \begin{pmatrix} \texttt{chrysler} & 1 \\ \texttt{usa} & 4 \\ \texttt{cat} & 3 \\ \texttt{dog} & 7 \\ \texttt{mouse} & 5 \end{pmatrix}$$

$$\mathbf{d}' = \begin{pmatrix} \texttt{chrysler} & 0.05 \\ \texttt{usa} & 0.2 \\ \texttt{cat} & 0.15 \\ \texttt{dog} & 0.35 \\ \texttt{mouse} & 0.25 \end{pmatrix}, \qquad \mathbf{q}' = \begin{pmatrix} \texttt{chrysler} & 0.2 \\ \texttt{usa} & 0.2 \\ \texttt{cat} & 0.2 \\ \texttt{dog} & 0.2 \\ \texttt{elephant} & 0.2 \end{pmatrix}$$

Vector Space Model

Example

$$\mathbf{d} = \begin{pmatrix} \texttt{chrysler} & w_1 \\ \texttt{usa} & w_2 \\ \texttt{cat} & w_3 \\ \texttt{dog} & w_4 \\ \texttt{mouse} & w_5 \end{pmatrix} = \begin{pmatrix} \texttt{chrysler} & 1 \\ \texttt{usa} & 4 \\ \texttt{cat} & 3 \\ \texttt{dog} & 7 \\ \texttt{mouse} & 5 \end{pmatrix}$$

$$\mathbf{d}' = \begin{pmatrix} \texttt{chrysler} & 0.05 \\ \texttt{usa} & 0.2 \\ \texttt{cat} & 0.15 \\ \texttt{dog} & 0.35 \\ \texttt{mouse} & 0.25 \\ \texttt{elephant} & 0.0 \end{pmatrix}, \qquad \mathbf{q}' = \begin{pmatrix} \texttt{chrysler} & 0.2 \\ \texttt{usa} & 0.2 \\ \texttt{cat} & 0.2 \\ \texttt{dog} & 0.2 \\ \texttt{mouse} & 0.0 \\ \texttt{elephant} & 0.2 \end{pmatrix}$$

The angle φ between $\mathbf{d}'$ and $\mathbf{q}'$ is about $48°$, $\cos(\varphi) \approx 0.67$.

The weights in d' and q' denote the relative term frequency $w'_i = \frac{w_i}{\sum_{j=1}^{5} w_j}$. Dimensions are aligned with zero padding. The product $\mathbf{d}'^T\mathbf{q}' = 0.15$, the norms $||\mathbf{d}'|| = 0.5$ and $||\mathbf{q}'|| = 0.447$.

Vector Space Model

Term Weighting: *tf* · *idf* [BIM Relevance Function]

To compute the weight w for a term t from document d under the vector space model, the most commonly employed term weighting scheme $\omega(t)$ is *tf* · *idf*:

- *tf*(t, d) denotes the normalized *term frequency* of term t in document d.
  The basic idea is that the importance of term t is proportional to its frequency in document d. However, t's importance does not increase linearly: the raw frequency must be normalized.
- *df*(t, D) denotes the *document frequency* of term t in document collection D.
  It counts the number of documents that contain t at least once.
- *idf*(t, D) denotes the *inverse document frequency*:

$$\mathit{idf}(t, D) = \log \frac{|D|}{\mathit{df}(t, D)}$$

  The importance of term t in general is inversely proportional to its document frequency.

A term weight ω for term t in document $d \in D$ is computed as follows:

$$\omega(t) = \mathit{tf}(t, d) \cdot \mathit{idf}(t, D).$$

Vector Space Model

Term Weighting: *tf* · *idf* [BIM Relevance Function]

To compute the weight w for a term t from document d under the vector space model, the most commonly employed term weighting scheme $\omega(t)$ is *tf* · *idf*:

- *tf*(t, d) denotes the normalized *term frequency* of term t in document d.
  The basic idea is that the importance of term t is proportional to its frequency in document d. However, t's importance does not increase linearly: the raw frequency must be normalized.

- *df*(t, D) denotes the *document frequency* of term t in document collection D.
  It counts the number of documents that contain t at least once.

- *idf*(t, D) denotes the *inverse document frequency*:

$$\mathit{idf}(t, D) = \log \frac{|D|}{\mathit{df}(t, D)}$$

  The importance of term t in general is inversely proportional to its document frequency.

A term weight ω for term t in document $d \in D$ is computed as follows:

$$\omega(t) = \mathit{tf}(t, d) \cdot \mathit{idf}(t, D).$$

Vector Space Model

Term Weighting: *tf* · *idf* [BIM Relevance Function]

To compute the weight w for a term t from document d under the vector space model, the most commonly employed term weighting scheme $\omega(t)$ is *tf* · *idf*:

- *tf*(t, d) denotes the normalized *term frequency* of term t in document d.
  The basic idea is that the importance of term t is proportional to its frequency in document d. However, t's importance does not increase linearly: the raw frequency must be normalized.
- *df*(t, D) denotes the *document frequency* of term t in document collection D.
  It counts the number of documents that contain t at least once.
- *idf*(t, D) denotes the *inverse document frequency*:

$$\textit{idf}(t, D) = \log \frac{|D|}{\textit{df}(t, D)}$$

  The importance of term t in general is inversely proportional to its document frequency.

A term weight ω for term t in document $d \in D$ is computed as follows:

$$\omega(t) = \textit{tf}(t, d) \cdot \textit{idf}(t, D).$$

Vector Space Model

Term Weighting: *tf* · *idf* [BIM Relevance Function]

To compute the weight w for a term t from document d under the vector space model, the most commonly employed term weighting scheme $\omega(t)$ is *tf* · *idf*:

- *tf*(t, d) denotes the normalized *term frequency* of term t in document d.
  The basic idea is that the importance of term t is proportional to its frequency in document d. However, t's importance does not increase linearly: the raw frequency must be normalized.

- *df*(t, D) denotes the *document frequency* of term t in document collection D.
  It counts the number of documents that contain t at least once.

- *idf*(t, D) denotes the *inverse document frequency*:

$$\textit{idf}(t, D) = \log \frac{|D|}{\textit{df}(t, D)}$$

  The importance of term t in general is inversely proportional to its document frequency.

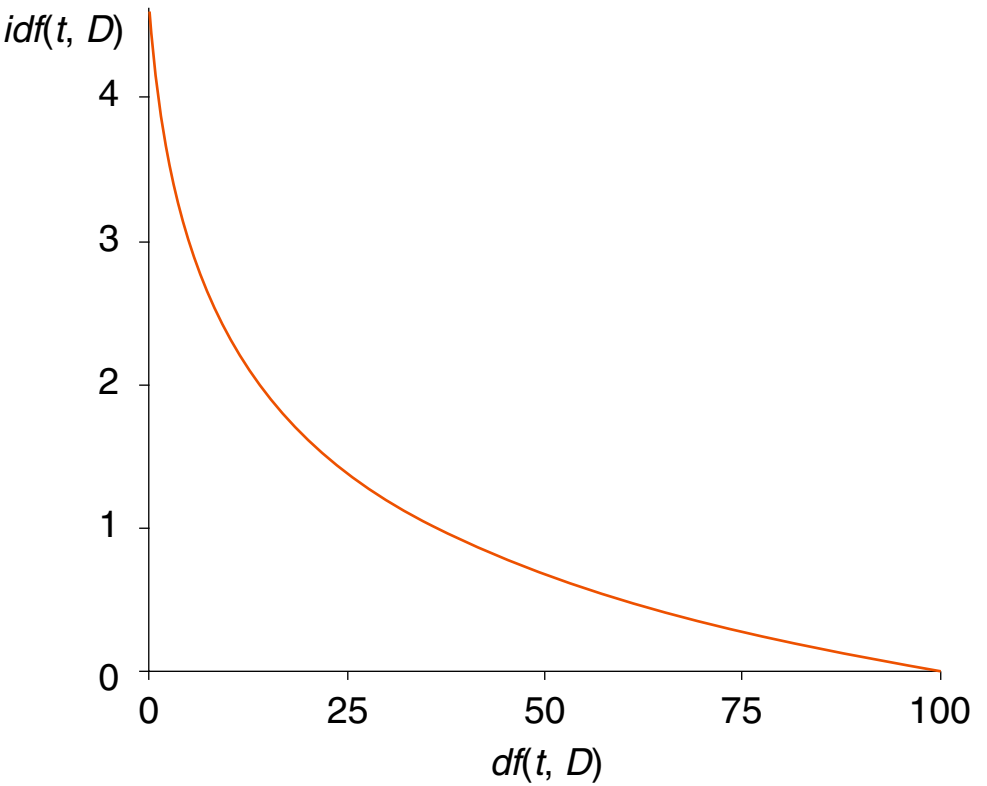
A term weight ω for term t in document $d \in D$ is computed as follows:

$$\omega(t) = \textit{tf}(t, d) \cdot \textit{idf}(t, D).$$

Vector Space Model

Term Weighting: $tf \cdot idf$

Plot of the function $idf(t, D) = \log \frac{|D|}{df(t, D)}$ for $|D| = 100$.



Remarks:

- Term frequency weighting was invented by Hans Peter Luhn: “There is also the probability that the more frequently a notion and combination of notions occur, the more importance the author attaches to them as reflecting the essence of his overall idea.” [Luhn 1957]
- The importance of a term t for a document d is not linearly correlated with its frequency. Several normalization factors have been proposed [Wikipedia]:
  - $\mathit{tf}(t,d)/|d|$
  - $1+\log(\mathit{tf}(t,d))$ for $\mathit{tf}(t,d)>0$
  - $k+(1-k)\frac{\mathit{tf}(t,d)}{\max_{t'\in d}(\mathit{tf}(t',d))}$, where k serves as smoothing term; typically $k=0.4$
- Inverse document frequency weighting was invented by Karen Spärck Jones: “it seems we should treat matches on non-frequent terms as more valuable than ones on frequent terms, without disregarding the latter altogether. The natural solution is to correlate a term’s matching value with its collection frequency.” [Spärck Jones 1972]
- Spärck Jones gives little theoretical justification for her intuition. Given the success of *idf* in practice, over the decades, numerous attempts at a theoretical justification have been made. A comprehensive overview has been compiled by Robertson 2004.
- For example, interpreting the term $\frac{|D|}{\mathit{df}(t,D)}$ as inverse of the probability $P_{\mathit{df}}(t)=\frac{\mathit{df}(t,D)}{|D|}$ of t occurring in a random document in D yields $\mathit{idf}(t,D)=\log\frac{|D|}{\mathit{df}(t,D)}=-\log P_{\mathit{df}}(t)$. Logarithms fit relevance functions ρ since both are additive, yielding the interpretation: “The less likely (on a random basis) it is that a given combination of terms occurs, the more likely it is that a document containing this combination is relevant to the question.” [Robertson 1972]

Vector Space Model

Discussion

Advantages:

- Improved retrieval performance compared to Boolean retrieval
- Partial query matching: not all query terms need to be present in a document for it to be retrieved
- The relevance function ρ defines a ranking among the retrieved documents with respect to their computed similarity to the query

Disadvantages:

- Index terms are assumed to occur independent of one another

Chapter IR:III

III. Retrieval Models

- Overview of Retrieval Models
- Boolean Retrieval
- Vector Space Model
- Binary Independence Model
- Okapi BM25
- Divergence From Randomness
- Latent Semantic Indexing
- Explicit Semantic Analysis
- Language Models
- Combining Evidence
- Learning to Rank

Binary Independence Model

Retrieval Model $\mathcal{R} = \langle \mathbf{D}, \mathbf{Q}, \rho \rangle$ [Generic Model] [Boolean] [VSM] [BIM] [BM25] [LSI] [ESA] [LM]

Document representations $\mathbf{D}$.

- $T = \{t_1, \ldots, t_m\}$ is the set of m index terms (stemmed words).
- $\mathbf{d} = \{\mathbf{d}(t_1), \ldots, \mathbf{d}(t_m)\}$ is a set of random variables over d and T.
- $\mathbf{d}(t) = 1$ if t occurs in d, and $\mathbf{d}(t) = 0$ otherwise.

Query representations $\mathbf{Q}$.

- $\mathbf{q} = \{\mathbf{q}(t_1), \ldots, \mathbf{q}(t_m)\}$ is a set of random variables over q and T.
- $\mathbf{q}(t) = 1$ if t occurs in d, and $\mathbf{q}(t) = 0$ otherwise.

Relevance function ρ. [Probability ranking principle]

- $\mathit{rel} : D \times Q \rightarrow \{0, 1\}$ indicates the true relevance of d to q for a given user.
- $r = \mathit{rel}(d, q)$ is a random variable indicating user relevance for a given d and q.
- $\rho(d, q) = P(r = 1 \mid \mathbf{d}, \mathbf{q})$, the probability of relevance of d for q for the user.

Binary Independence Model

Retrieval Model $\mathcal{R} = \langle \mathbf{D}, \mathbf{Q}, \rho \rangle$ [Generic Model] [Boolean] [VSM] [BIM] [BM25] [LSI] [ESA] [LM]

Document representations $\mathbf{D}$.

- $T = \{t_1, \ldots, t_m\}$ is the set of m index terms (stemmed words).
- $\mathbf{d} = \{\mathbf{d}(t_1), \ldots, \mathbf{d}(t_m)\}$ is a set of random variables over d and T.
- $\mathbf{d}(t) = 1$ if t occurs in d, and $\mathbf{d}(t) = 0$ otherwise.

Query representations $\mathbf{Q}$.

- $\mathbf{q} = \{\mathbf{q}(t_1), \ldots, \mathbf{q}(t_m)\}$ is a set of random variables over q and T.
- $\mathbf{q}(t) = 1$ if t occurs in d, and $\mathbf{q}(t) = 0$ otherwise.

Relevance function ρ. [Probability ranking principle]

- $\mathit{rel} : D \times Q \to \{0, 1\}$ indicates the true relevance of d to q for a given user.
- $r = \mathit{rel}(d, q)$ is a random variable indicating user relevance for a given d and q.
- $\rho(d, q) = P(r = 1 \mid \mathbf{d}, \mathbf{q})$, the probability of relevance of d for q for the user.

Remarks:

- The model is also known as Okapi model (based on the Okapi Information Retrieval System), City model (based on its origin, the City University, London), or simply as the probabilistic model.
- The joint probability space $(\Omega, \mathcal{P}(\Omega), P)$ underlying the binary independence model is given by the sample space $\Omega = \{0, 1\} \times \mathcal{P}(T)$, where $\mathcal{P}(T)$ denotes the set of all binary document vectors over the set of terms T.

Binary Independence Model

Relevance Function ρ: Derivation

Let $r = \mathit{rel}(d, q)$ denote the true (binary) relevance, and let $\mathbf{d}, \mathbf{q}$ represent document d and query q:

$$P(r = 1 \mid \mathbf{d}, \mathbf{q}) \overset{\text{rank}}{=} \frac{P(r = 1 \mid \mathbf{d}, \mathbf{q})}{P(r = 0 \mid \mathbf{d}, \mathbf{q})} \tag{1}$$

(1) Rank-preserving replacement of $P(A)$ by the odds $\frac{P(A)}{P(\overline{A})}$ in favor of event A.

Binary Independence Model

Relevance Function ρ: Derivation

Let $r = \mathit{rel}(d, q)$ denote the true (binary) relevance, and let $\mathbf{d}, \mathbf{q}$ represent document d and query q:

$$P(r = 1 \mid \mathbf{d}, \mathbf{q}) \overset{\text{rank}}{=} \frac{P(r = 1 \mid \mathbf{d}, \mathbf{q})}{P(r = 0 \mid \mathbf{d}, \mathbf{q})} \tag{1}$$

$$= \frac{P(\mathbf{d} \mid r = 1, \mathbf{q})\ P(r = 1 \mid \mathbf{q})}{P(\mathbf{d} \mid r = 0, \mathbf{q})\ P(r = 0 \mid \mathbf{q})} \tag{2}$$

(1) Rank-preserving replacement of $P(A)$ by the odds $\frac{P(A)}{P(\overline{A})}$ in favor of event A.

(2) Application of Bayes' rule. The common denominator $P(\mathbf{d} \mid \mathbf{q})$ is canceled.

Binary Independence Model

Relevance Function ρ: Derivation

Let $r = \mathit{rel}(d, q)$ denote the true (binary) relevance, and let $\mathbf{d}, \mathbf{q}$ represent document d and query q:

$$P(r = 1 \mid \mathbf{d}, \mathbf{q}) \overset{\text{rank}}{=} \frac{P(r = 1 \mid \mathbf{d}, \mathbf{q})}{P(r = 0 \mid \mathbf{d}, \mathbf{q})} \tag{1}$$

$$= \frac{P(\mathbf{d} \mid r = 1, \mathbf{q})\ P(r = 1 \mid \mathbf{q})}{P(\mathbf{d} \mid r = 0, \mathbf{q})\ P(r = 0 \mid \mathbf{q})} \tag{2}$$

$$\overset{\text{rank}}{=} \frac{P(\mathbf{d} \mid r = 1, \mathbf{q})}{P(\mathbf{d} \mid r = 0, \mathbf{q})} \tag{3}$$

(1) Rank-preserving replacement of $P(A)$ by the odds $\frac{P(A)}{P(\overline{A})}$ in favor of event A.

(2) Application of Bayes' rule. The common denominator $P(\mathbf{d} \mid \mathbf{q})$ is canceled.

(3) Rank-preserving omission of $\frac{P(r = 1 \mid \mathbf{q})}{P(r = 0 \mid \mathbf{q})}$; it does not depend on $\mathbf{d}$.

Binary Independence Model

Relevance Function ρ: Derivation

Let $\mathbf{d}(t)$ denote if t occurs in d.

$$\frac{P(\mathbf{d} \mid r = 1, \mathbf{q})}{P(\mathbf{d} \mid r = 0, \mathbf{q})} = \prod_{t \in T} \frac{P(\mathbf{d}(t) \mid r = 1, \mathbf{q})}{P(\mathbf{d}(t) \mid r = 0, \mathbf{q})} \tag{4}$$

(4) Assuming independence between terms.

Binary Independence Model

Relevance Function ρ: Derivation

Let $\mathbf{d}(t)$ denote if t occurs in d.

$$\frac{P(\mathbf{d} \mid r=1,\mathbf{q})}{P(\mathbf{d} \mid r=0,\mathbf{q})} = \prod_{t \in T} \frac{P(\mathbf{d}(t) \mid r=1,\mathbf{q})}{P(\mathbf{d}(t) \mid r=0,\mathbf{q})} \tag{4}$$

$$= \prod_{t \in \mathbf{d}} \frac{P(\mathbf{d}(t)=1 \mid r=1,\mathbf{q})}{P(\mathbf{d}(t)=1 \mid r=0,\mathbf{q})} \prod_{t \notin \mathbf{d}} \frac{P(\mathbf{d}(t)=0 \mid r=1,\mathbf{q})}{P(\mathbf{d}(t)=0 \mid r=0,\mathbf{q})} \tag{5}$$

(4) Assuming independence between terms.

(5) Separation of the the two possible cases for $\mathbf{d}(t)$,
where $t \in \mathbf{d}$ means $t \in T : \mathbf{d}(t) = 1$ and $t \notin \mathbf{d}$ means $t \in T : \mathbf{d}(t) = 0$.

Binary Independence Model

Relevance Function ρ: Derivation

Let $\mathbf{d}(t)$ denote if t occurs in d.

$$\frac{P(\mathbf{d} \mid r=1,\mathbf{q})}{P(\mathbf{d} \mid r=0,\mathbf{q})} = \prod_{t \in T} \frac{P(\mathbf{d}(t) \mid r=1,\mathbf{q})}{P(\mathbf{d}(t) \mid r=0,\mathbf{q})} \tag{4}$$

$$= \prod_{t \in \mathbf{d}} \frac{P(\mathbf{d}(t)=1 \mid r=1,\mathbf{q})}{P(\mathbf{d}(t)=1 \mid r=0,\mathbf{q})} \prod_{t \notin \mathbf{d}} \frac{P(\mathbf{d}(t)=0 \mid r=1,\mathbf{q})}{P(\mathbf{d}(t)=0 \mid r=0,\mathbf{q})} \tag{5}$$

$$= \prod_{t \in \mathbf{d}} \frac{p_t}{s_t} \prod_{t \notin \mathbf{d}} \frac{1-p_t}{1-s_t} \tag{6}$$

(4) Assuming independence between terms.

(5) Separation of the the two possible cases for $\mathbf{d}(t)$,
where $t \in \mathbf{d}$ means $t \in T : \mathbf{d}(t) = 1$ and $t \notin \mathbf{d}$ means $t \in T : \mathbf{d}(t) = 0$.

(6) Abbreviation: $p_t = P(\mathbf{d}(t)=1 \mid r=1,\mathbf{q})$ and $s_t = P(\mathbf{d}(t)=1 \mid r=0,\mathbf{q})$.

Binary Independence Model

Relevance Function ρ: Derivation

$$\prod_{t \in \mathbf{d}} \frac{p_t}{s_t} \prod_{t \notin \mathbf{d}} \frac{1-p_t}{1-s_t} \overset{\text{rank}}{=} \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t}{s_t} \prod_{\substack{t \in \mathbf{q}: \\ t \notin \mathbf{d}}} \frac{1-p_t}{1-s_t} \tag{7}$$

(7) Assumption: $p_t = s_t$ for $t \notin \mathbf{q}$ (i.e., non-query terms are equally likely in relevant and non-relevant documents). Rank-preserving omission of these factors.

Binary Independence Model

Relevance Function ρ: Derivation

$$\prod_{t \in \mathbf{d}} \frac{p_t}{s_t} \prod_{t \notin \mathbf{d}} \frac{1-p_t}{1-s_t} \overset{\text{rank}}{=} \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t}{s_t} \prod_{\substack{t \in \mathbf{q}: \\ t \notin \mathbf{d}}} \frac{1-p_t}{1-s_t} \tag{7}$$

$$= \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t}{s_t} \prod_{\substack{t \in \mathbf{q}: \\ t \notin \mathbf{d}}} \frac{1-p_t}{1-s_t} \tag{8}$$

(7) Assumption: $p_t = s_t$ for $t \notin \mathbf{q}$ (i.e., non-query terms are equally likely in relevant and non-relevant documents). Rank-preserving omission of these factors.

(8) Addition of all missing query terms to the right product and division by the added factors to fulfill the equation.

Binary Independence Model

Relevance Function ρ: Derivation

$$\prod_{t \in \mathbf{d}} \frac{p_t}{s_t} \prod_{t \notin \mathbf{d}} \frac{1-p_t}{1-s_t} \overset{\text{rank}}{=} \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t}{s_t} \prod_{\substack{t \in \mathbf{q}: \\ t \notin \mathbf{d}}} \frac{1-p_t}{1-s_t} \tag{7}$$

$$= \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t}{s_t} \prod_{t \in \mathbf{q}} \frac{1-p_t}{1-s_t} \Big/ \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{1-p_t}{1-s_t} \tag{8}$$

(7) Assumption: $p_t = s_t$ for $t \notin \mathbf{q}$ (i.e., non-query terms are equally likely in relevant and non-relevant documents). Rank-preserving omission of these factors.

(8) Addition of all missing query terms to the right product and division by the added factors to fulfill the equation.

Binary Independence Model

Relevance Function ρ: Derivation

$$\prod_{t \in \mathbf{d}} \frac{p_t}{s_t} \prod_{t \notin \mathbf{d}} \frac{1-p_t}{1-s_t} \overset{\text{rank}}{=} \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t}{s_t} \prod_{\substack{t \in \mathbf{q}: \\ t \notin \mathbf{d}}} \frac{1-p_t}{1-s_t} \tag{7}$$

$$= \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t}{s_t} \prod_{t \in \mathbf{q}} \frac{1-p_t}{1-s_t} \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{1-s_t}{1-p_t} \tag{8}$$

(7) Assumption: $p_t = s_t$ for $t \notin \mathbf{q}$ (i.e., non-query terms are equally likely in relevant and non-relevant documents). Rank-preserving omission of these factors.

(8) Addition of all missing query terms to the right product and division by the added factors to fulfill the equation.

Binary Independence Model

Relevance Function ρ: Derivation

$$\prod_{t \in \mathbf{d}} \frac{p_t}{s_t} \prod_{t \notin \mathbf{d}} \frac{1-p_t}{1-s_t} \overset{\text{rank}}{=} \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t}{s_t} \prod_{\substack{t \in \mathbf{q}: \\ t \notin \mathbf{d}}} \frac{1-p_t}{1-s_t} \tag{7}$$

$$= \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t}{s_t} \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{1-s_t}{1-p_t} \prod_{t \in \mathbf{q}} \frac{1-p_t}{1-s_t} \tag{8}$$

(7) Assumption: $p_t = s_t$ for $t \notin \mathbf{q}$ (i.e., non-query terms are equally likely in relevant and non-relevant documents). Rank-preserving omission of these factors.

(8) Addition of all missing query terms to the right product and division by the added factors to fulfill the equation.

Binary Independence Model

Relevance Function ρ: Derivation

$$\prod_{t \in \mathbf{d}} \frac{p_t}{s_t} \prod_{t \notin \mathbf{d}} \frac{1-p_t}{1-s_t} \overset{\text{rank}}{=} \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t}{s_t} \prod_{\substack{t \in \mathbf{q}: \\ t \notin \mathbf{d}}} \frac{1-p_t}{1-s_t} \tag{7}$$

$$= \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t(1-s_t)}{s_t(1-p_t)} \prod_{t \in \mathbf{q}} \frac{1-p_t}{1-s_t} \tag{8}$$

(7) Assumption: $p_t = s_t$ for $t \notin \mathbf{q}$ (i.e., non-query terms are equally likely in relevant and non-relevant documents). Rank-preserving omission of these factors.

(8) Addition of all missing query terms to the right product and division by the added factors to fulfill the equation.

Binary Independence Model

Relevance Function ρ: Derivation

$$\prod_{t \in \mathbf{d}} \frac{p_t}{s_t} \prod_{t \notin \mathbf{d}} \frac{1-p_t}{1-s_t} \overset{\text{rank}}{=} \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t}{s_t} \prod_{\substack{t \in \mathbf{q}: \\ t \notin \mathbf{d}}} \frac{1-p_t}{1-s_t} \tag{7}$$

$$= \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t(1-s_t)}{s_t(1-p_t)} \prod_{t \in \mathbf{q}} \frac{1-p_t}{1-s_t} \tag{8}$$

$$\overset{\text{rank}}{=} \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t(1-s_t)}{s_t(1-p_t)} \tag{9}$$

(7) Assumption: $p_t = s_t$ for $t \notin \mathbf{q}$ (i.e., non-query terms are equally likely in relevant and non-relevant documents). Rank-preserving omission of these factors.

(8) Addition of all missing query terms to the right product and division by the added factors to fulfill the equation.

(9) Rank-preserving omission of the right product; it does not depend on $\mathbf{d}$.

Binary Independence Model

Relevance Function ρ: Derivation

$$\prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t(1-s_t)}{s_t(1-p_t)} \quad \overset{\text{rank}}{=} \quad \log \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t(1-s_t)}{s_t(1-p_t)} \tag{10}$$

(10) Rank-preserving logarithmization to allow for computations that do not underflow common floating point number formats.

Binary Independence Model

Relevance Function ρ: Derivation

$$\prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t(1-s_t)}{s_t(1-p_t)} \overset{\text{rank}}{=} \log \prod_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \frac{p_t(1-s_t)}{s_t(1-p_t)} = \sum_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \underbrace{\log \frac{p_t(1-s_t)}{s_t(1-p_t)}}_{:= \omega_{\text{RSJ}}} \tag{10}$$

(10) Rank-preserving logarithmization to allow for computations that do not underflow common floating point number formats.

In effect, we accumulate for each term $t \in \mathbf{q}$ the log odds ratio of the odds in favor and the odds against t occurring in $\mathbf{d}$ if the document d is (non-)relevant to query q.

This ratio tells us how much more likely it is that t occurs in $\mathbf{d}$ if d is relevant to q.

RSJ $\sim$ Robertson Spärck-Jones

Binary Independence Model

Relevance Function ρ: Estimation

Let D denote the document collection and D_t the subset containing term t.

$$\sum_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \log \frac{p_t(1-s_t)}{s_t(1-p_t)} \quad \overset{\text{rank}}{=} \quad \sum_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \log \frac{1-s_t}{s_t} \tag{11}$$

(11) Assumption that a term $t \in \mathbf{q}$ is equally likely to be present or absent in a random relevant document: $p_t = 0.5$. This cancels p_t and $1 - p_t$.

Binary Independence Model

Relevance Function ρ: Estimation

Let D denote the document collection and D_t the subset containing term t.

$$\sum_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \log \frac{p_t(1-s_t)}{s_t(1-p_t)} \quad \overset{\text{rank}}{=} \quad \sum_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \log \frac{1-s_t}{s_t} \tag{11}$$

$$\overset{\text{rank}}{=} \sum_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \log \frac{|D| - |D_t| + 0.5}{|D_t| + 0.5} \tag{12}$$

(11) Assumption that a term $t \in \mathbf{q}$ is equally likely to be present or absent in a random relevant document: $p_t = 0.5$. This cancels p_t and $1 - p_t$.

(12) Maximum likelihood estimation of $s_t = P(\mathbf{d}(t) = 1 \mid r = 0, \mathbf{q})$:

$$s_t = \frac{|D_t| + 0.5}{|D| + 1.0} \,,$$

where adding 0.5 (1.0) is used for smoothing (avoiding zeros). Assumption that $|D|$ represents the set of non-relevant documents.

Remarks:

- Adding 0.5 (1.0) in this way is a simple form of smoothing. For trials with categorical outcomes (such as noting the presence or absence of a term), one way to estimate the probability of an event from data is simply to count the number of times an event occurred divided by the total number of trials. This relative is referred to as the relative frequency of the event. Estimating the probability as the relative frequency is the maximum likelihood estimate (or MLE), maximum because this value makes the observed data maximally likely. However, if we simply use the MLE, then the probability given to events we happened to see is usually too high, whereas other events may be completely unseen and giving them as a probability estimate their relative frequency of 0 is both an underestimate and normally breaks our models; anything multiplied by 0 is 0.

  Simultaneously decreasing the estimated probability of seen events and increasing the probability of unseen events is referred to as smoothing. One simple way of smoothing is to add a number α (β) to each of the observed counts (totals). These pseudocounts correspond to the use of a uniform distribution over the vocabulary as a Bayesian prior. We initially assume a uniform distribution over events, where the size of α denotes the strength of our belief in uniformity, and we then update the probability based on observed events. Because our belief in uniformity is weak, we use $\alpha = 0.5, \beta = 1.0$. This is a form of maximum a posteriori (MAP) estimation, where we choose the most likely point value for probabilities based on the prior and the observed evidence. [Manning/Raghavan/Schütze 2008]

Binary Independence Model

Relevance Function ρ: Example

$$q = (\mathbf{a}\ \ \mathbf{c}\ \ \mathbf{h})$$

$$\begin{aligned} D = \{\ & d_1 = (\mathbf{a}\ \mathtt{b}\ \mathbf{c}\ \mathtt{b}\ \mathtt{d}), \\ & d_2 = (\mathtt{b}\ \mathtt{e}\ \mathtt{f}\ \mathtt{b}), \\ & d_3 = (\mathtt{b}\ \mathtt{g}\ \mathbf{c}\ \mathtt{d}), \\ & d_4 = (\mathtt{b}\ \mathtt{d}\ \mathtt{e}), \\ & d_5 = (\mathbf{a}\ \mathtt{b}\ \mathtt{e}\ \mathtt{g}), \\ & d_6 = (\mathtt{b}\ \mathtt{g}\ \mathbf{h})\ \} \end{aligned}$$

Term t	**a**	b	**c**	d	e	f	g	**h**
$\lvert D_t \rvert$	2	6	2	3	3	1	3	1
s_t	0.4	0.9	0.4	0.5	0.5	0.2	0.5	0.2

$$\begin{aligned} \rho(\mathbf{d}_i, \mathbf{q}) &= \underbrace{\log \frac{1-0.4}{0.4}}_{t=\mathbf{a}} + \underbrace{\log \frac{1-0.4}{0.4}}_{t=\mathbf{c}} + \underbrace{\log \frac{1-0.2}{0.2}}_{t=\mathbf{h}} \\ &= \log 1.5 + \log 1.5 + \log 4 \\ &= 0.1761 + 0.1761 + 0.6021 \\ &= 0.9543 \end{aligned}$$

Document	$\rho(\mathbf{d}_i, \mathbf{q})$
d_1	
d_2	
d_3	
d_4	
d_5	
d_6	

Binary Independence Model

Relevance Function ρ: Example

$$q = (\mathbf{a}\ \ \mathbf{c}\ \ \mathbf{h})$$

$$\begin{aligned} D = \{\ & d_1 = (\mathbf{a}\ \mathtt{b}\ \mathbf{c}\ \mathtt{b}\ \mathtt{d}), \\ & d_2 = (\mathtt{b}\ \mathtt{e}\ \mathtt{f}\ \mathtt{b}), \\ & d_3 = (\mathtt{b}\ \mathtt{g}\ \mathbf{c}\ \mathtt{d}), \\ & d_4 = (\mathtt{b}\ \mathtt{d}\ \mathtt{e}), \\ & d_5 = (\mathbf{a}\ \mathtt{b}\ \mathtt{e}\ \mathtt{g}), \\ & d_6 = (\mathtt{b}\ \mathtt{g}\ \mathbf{h})\ \} \end{aligned}$$

Term t	**a**	b	**c**	d	e	f	g	**h**
$\lvert D_t \rvert$	2	6	2	3	3	1	3	1
s_t	0.4	0.9	0.4	0.5	0.5	0.2	0.5	0.2

$$\begin{aligned} \rho(\mathbf{d}_1, \mathbf{q}) &= \underbrace{\log \frac{1-0.4}{0.4}}_{t=\mathbf{a}} + \underbrace{\log \frac{1-0.4}{0.4}}_{t=\mathbf{c}} + \underbrace{\log \frac{1-0.2}{0.2}}_{t=\mathbf{h}} \\ &= \log 1.5 + \log 1.5 + \log 4 \\ &= 0.1761 + 0.1761 + 0.6021 \\ &= 0.3522 \end{aligned}$$

Document	$\rho(\mathbf{d}_i, \mathbf{q})$
d_1	0.3522
d_2	
d_3	
d_4	
d_5	
d_6	

Binary Independence Model

Relevance Function ρ: Example

$$q = (\mathbf{a}\ \mathbf{c}\ \mathbf{h})$$

Term t	**a**	b	**c**	d	e	f	g	**h**
$\lvert D_t \rvert$	2	6	2	3	3	1	3	1
s_t	0.4	0.9	0.4	0.5	0.5	0.2	0.5	0.2

$$\begin{aligned} D = \{\ & d_1 = (\mathbf{a}\ \mathtt{b}\ \mathbf{c}\ \mathtt{b}\ \mathtt{d}), \\ & d_2 = (\mathtt{b}\ \mathtt{e}\ \mathtt{f}\ \mathtt{b}), \\ & d_3 = (\mathtt{b}\ \mathtt{g}\ \mathbf{c}\ \mathtt{d}), \\ & d_4 = (\mathtt{b}\ \mathtt{d}\ \mathtt{e}), \\ & d_5 = (\mathbf{a}\ \mathtt{b}\ \mathtt{e}\ \mathtt{g}), \\ & d_6 = (\mathtt{b}\ \mathtt{g}\ \mathbf{h})\ \} \end{aligned}$$

$$\begin{aligned} \rho(\mathbf{d}_2, \mathbf{q}) &= \underbrace{\log \frac{1-0.4}{0.4}}_{t=\mathbf{a}} + \underbrace{\log \frac{1-0.4}{0.4}}_{t=\mathbf{c}} + \underbrace{\log \frac{1-0.2}{0.2}}_{t=\mathbf{h}} \\ &= \log 1.5 + \log 1.5 + \log 4 \\ &= 0.1761 + 0.1761 + 0.6021 \\ &= 0 \end{aligned}$$

Document	$\rho(\mathbf{d}_i, \mathbf{q})$
d_1	0.3522
d_2	0
d_3	
d_4	
d_5	
d_6	

Binary Independence Model

Relevance Function ρ: Example

$$q = (\mathbf{a}\ \ \mathbf{c}\ \ \mathbf{h})$$

$$\begin{aligned} D = \{\ & d_1 = (\mathbf{a}\ \mathtt{b}\ \mathbf{c}\ \mathtt{b}\ \mathtt{d}), \\ & d_2 = (\mathtt{b}\ \mathtt{e}\ \mathtt{f}\ \mathtt{b}), \\ & d_3 = (\mathtt{b}\ \mathtt{g}\ \mathbf{c}\ \mathtt{d}), \\ & d_4 = (\mathtt{b}\ \mathtt{d}\ \mathtt{e}), \\ & d_5 = (\mathbf{a}\ \mathtt{b}\ \mathtt{e}\ \mathtt{g}), \\ & d_6 = (\mathtt{b}\ \mathtt{g}\ \mathbf{h})\ \} \end{aligned}$$

Term t	**a**	b	**c**	d	e	f	g	**h**
$\lvert D_t \rvert$	2	6	2	3	3	1	3	1
s_t	0.4	0.9	0.4	0.5	0.5	0.2	0.5	0.2

$$\begin{aligned} \rho(\mathbf{d}_3, \mathbf{q}) &= \underbrace{\log \frac{1-0.4}{0.4}}_{t=\mathbf{a}} + \underbrace{\log \frac{1-0.4}{0.4}}_{t=\mathbf{c}} + \underbrace{\log \frac{1-0.2}{0.2}}_{t=\mathbf{h}} \\ &= \log 1.5 + \log 1.5 + \log 4 \\ &= 0.1761 + 0.1761 + 0.6021 \\ &= 0.1761 \end{aligned}$$

Document	$\rho(\mathbf{d}_i, \mathbf{q})$
d_1	0.3522
d_2	0
d_3	0.1761
d_4	
d_5	
d_6	

Binary Independence Model

Relevance Function ρ: Example

$$q = (\mathbf{a}\ \mathbf{c}\ \mathbf{h})$$

$$D = \{\ d_1 = (\mathbf{a}\ \mathtt{b}\ \mathbf{c}\ \mathtt{b}\ \mathtt{d}),\ d_2 = (\mathtt{b}\ \mathtt{e}\ \mathtt{f}\ \mathtt{b}),\ d_3 = (\mathtt{b}\ \mathtt{g}\ \mathbf{c}\ \mathtt{d}),\ d_4 = (\mathtt{b}\ \mathtt{d}\ \mathtt{e}),\ d_5 = (\mathbf{a}\ \mathtt{b}\ \mathtt{e}\ \mathtt{g}),\ d_6 = (\mathtt{b}\ \mathtt{g}\ \mathbf{h})\ \}$$

Term t	**a**	b	**c**	d	e	f	g	**h**
$\lvert D_t \rvert$	2	6	2	3	3	1	3	1
s_t	0.4	0.9	0.4	0.5	0.5	0.2	0.5	0.2

$$\begin{aligned}
\rho(\mathbf{d}_6, \mathbf{q}) &= \underbrace{\log \frac{1-0.4}{0.4}}_{t=\mathbf{a}} + \underbrace{\log \frac{1-0.4}{0.4}}_{t=\mathbf{c}} + \underbrace{\log \frac{1-0.2}{0.2}}_{t=\mathbf{h}} \\
&= \log 1.5 + \log 1.5 + \log 4 \\
&= 0.1761 + 0.1761 + 0.6021 \\
&= 0.6021
\end{aligned}$$

Document	$\rho(\mathbf{d}_i, \mathbf{q})$
d_1	0.3522
d_2	0
d_3	0.1761
d_4	0
d_5	0.1761
d_6	0.6021

Binary Independence Model

Relevance Function ρ: Example

$$q = (\mathbf{a}\ \ \mathbf{c}\ \ \mathbf{h})$$

Term t	**a**	b	**c**	d	e	f	g	**h**
$\lvert D_t \rvert$	2	6	2	3	3	1	3	1
s_t	0.4	0.9	0.4	0.5	0.5	0.2	0.5	0.2

$$\begin{aligned} D = \{\ & d_1 = (\mathbf{a}\ \mathtt{b}\ \mathbf{c}\ \mathtt{b}\ \mathtt{d}), \\ & d_2 = (\mathtt{b}\ \mathtt{e}\ \mathtt{f}\ \mathtt{b}), \\ & d_3 = (\mathtt{b}\ \mathtt{g}\ \mathbf{c}\ \mathtt{d}), \\ & d_4 = (\mathtt{b}\ \mathtt{d}\ \mathtt{e}), \\ & d_5 = (\mathbf{a}\ \mathtt{b}\ \mathtt{e}\ \mathtt{g}), \\ & d_6 = (\mathtt{b}\ \mathtt{g}\ \mathbf{h})\ \} \end{aligned}$$

$$\begin{aligned} \rho(\mathbf{d}_6, \mathbf{q}) &= \underbrace{\log \frac{1-0.4}{0.4}}_{t=\mathbf{a}} + \underbrace{\log \frac{1-0.4}{0.4}}_{t=\mathbf{c}} + \underbrace{\log \frac{1-0.2}{0.2}}_{t=\mathbf{h}} \\ &= \log 1.5 + \log 1.5 + \log 4 \\ &= 0.1761 + 0.1761 + 0.6021 \\ &= 0.6021 \end{aligned}$$

Ranking	$\rho(\mathbf{d}_i, \mathbf{q})$
d_6	0.6021
d_1	0.3522
d_3	0.1761
d_5	0.1761
d_2	0
d_4	0

Binary Independence Model

Relevance Function ρ: Example

$$q = (\mathbf{a}\ \mathbf{c}\ \mathbf{h})$$

Term t	**a**	b	**c**	d	e	f	g	**h**
$\lvert D_t \rvert$	2	6	2	3	3	1	3	1
s_t	0.4	0.9	0.4	0.5	0.5	0.2	0.5	0.2

$$\begin{aligned} D = \{\ & d_1 = (\mathbf{a}\ \mathtt{b}\ \mathbf{c}\ \mathtt{b}\ \mathtt{d}), \\ & d_2 = (\mathtt{b}\ \mathtt{e}\ \mathtt{f}\ \mathtt{b}), \\ & d_3 = (\mathtt{b}\ \mathtt{g}\ \mathbf{c}\ \mathtt{d}), \\ & d_4 = (\mathtt{b}\ \mathtt{d}\ \mathtt{e}), \\ & d_5 = (\mathbf{a}\ \mathtt{b}\ \mathtt{e}\ \mathtt{g}), \\ & d_6 = (\mathtt{b}\ \mathtt{g}\ \mathbf{h})\ \} \end{aligned}$$

$$\begin{aligned} \rho(\mathbf{d}_6, \mathbf{q}) &= \underbrace{\log \frac{1-0.4}{0.4}}_{t=\mathbf{a}} + \underbrace{\log \frac{1-0.4}{0.4}}_{t=\mathbf{c}} + \underbrace{\log \frac{1-0.2}{0.2}}_{t=\mathbf{h}} \\ &= \log 1.5 + \log 1.5 + \log 4 \\ &= 0.1761 + 0.1761 + 0.6021 \\ &= 0.6021 \end{aligned}$$

Ranking	$\rho(\mathbf{d}_i, \mathbf{q})$
d_6	0.6021
d_1	0.3522
d_3	0.1761
d_5	0.1761
d_2	0
d_4	0

Why is d_6 the most relevant document?

Binary Independence Model

Relevance Function ρ: Summary [Inverse Document Frequency]

$$\rho(\mathbf{d},\mathbf{q}) \quad = \quad P(r=1 \mid \mathbf{d},\mathbf{q}) \quad \propto \quad \sum_{\substack{t\in\mathbf{q}:\\ t\in\mathbf{d}}} \underbrace{\log \frac{|D| - |D_t| + 0.5}{|D_t| + 0.5}}_{:= \omega_{\text{BIM}} \approx \mathit{idf}(t,D)}$$

Assumptions:

1. Binary relevance of a document d to a query q, independent of all other documents.
2. Boolean representations $\mathbf{d}, \mathbf{q}$ of document d and query q.
3. Independence of word occurrence in documents.
4. Terms not in query $\mathbf{q}$ are equally likely to occur in relevant and non-relevant documents.
5. Terms in $\mathbf{q}$ are equally likely to occur or not to occur in a relevant document d.
6. The set of non-relevant documents is represented by the entire collection.

Remarks:

- It is a personal observation that almost every mathematically inclined graduate student in Information Retrieval attempts to formulate some sort of a non-independent model of IR within the first two to three years of his or her studies. The vast majority of these attempts yield no improvements and remain unpublished. [...] It is natural to wonder why this is the case – the classical model contains an obviously incorrect assumption about the language, and yet most attempts to relax that assumption produce no consistent improvements whatsoever.

  Contrary to popular belief, word independence is not a necessary assumption in the classical probabilistic model of IR. A necessary and sufficient condition is proportional interdependence [...]: *on average*, all the words in a given document have about as much interdependence under the relevant class as they do under the non-relevant class. [...] the only requirement is that whatever disbalance exists be constant across all documents. If there is anything wrong with the classical model, it is not independence but the assumptions made in the estimation process. [Lavrenko 2009]

Binary Independence Model

Discussion

Advantages:

- grounded in probabilistic theory
- performs well given some relevance feedback
- supplies theoretical justification of inverse document frequency

Disadvantages:

- in the absence of relevance feedback, only about 50% recall compared to a $tf \cdot idf$-based vector space model
- does not exploit term frequencies
- assumptions and rank-preserving simplifications do not generalize to retrieval scenarios other than ad hoc retrieval

Okapi BM25

Retrieval Model $\mathcal{R} = \langle \mathbf{D}, \mathbf{Q}, \rho \rangle$ [Generic Model] [Boolean] [VSM] [BIM] [BM25] [LSI] [ESA] [LM]

Document representations $\mathbf{D}$.

- $T = \{t_1, \ldots, t_m\}$ is the set of m index terms (word stems, without stop words).
- T is interpreted as set of dimensions of an m-dimensional vector space.
- $\omega : \mathbf{D} \times T \to \mathbf{R}$ is a term weighting function, quantifying term importance.
- $\mathbf{d} = (w_1, \ldots, w_m)^T$, where $w_i = \omega(\mathbf{d}, t_i)$ is the term weight of the i-th term in T.

Query representations $\mathbf{Q}$.

- $\mathbf{q} = (w_1, \ldots, w_m)^T$, where $w_i = \omega(\mathbf{q}, t_i)$ is the term weight of the i-th term in T.

Relevance function ρ.

- $\rho(d, q) = \sum_{t \in \mathbf{q}} \omega_{\text{BM25}}(t, \mathbf{d}, D)$ is the sum of BM25 term weights in $\mathbf{q}$ given $\mathbf{d}$.

Okapi BM25

Background

Empirical evidence suggests that term frequency is an important factor in determining the relevance of a document.

Relaxation of BIM to term frequencies within the 2-Poisson model:

- Change of the joint sample space to $\Omega = \{0,1\} \times \mathbf{N}^{|T|}$, where $\mathbf{N}^{|T|}$ denotes the set of document vectors with term frequency weights over the set of terms T.
- Starting point is the RSJ weight of the binary independence model:

$$P(r = 1 \mid \mathbf{d}, \mathbf{q}) \quad \propto \quad \omega_{\text{RSJ}} \quad = \quad \log \frac{p_t(1 - s_t)}{s_t(1 - p_t)}$$

- Estimation of $p_t = P(\mathbf{d}(t) = \mathit{tf}(t, \mathbf{d}) \mid r = 1, \mathbf{q})$ as mixture of two Poisson distributions, distinguishing "elite" terms that occur unusually frequently from others. An elite term t encodes whether d is about the concept underlying t.
- Problems: Poisson distribution is a poor fit; too many parameters
- Approach: empirical approximation of the term weight ω_{RSJ}
- Resulted in the successful term weighting scheme Okapi BM25.

Remarks:

- "Okapi" is the name of a retrieval system developed at City University London. "BM" stands for Best Match, and the number 25 refers to the best-performing variant tried. Other variants include BM0, BM1, BM11, and BM15. Here is an overview of all variants.
- We assume that each document is generated by filling a certain number of word-positions (fixed length) from a vocabulary of words. Furthermore, we assume a simple multinomial distribution over words, so that for each position each word has a fixed (small) probability of being chosen, independent of what other words have been chosen for other positions. Then it follows that the distribution of *tf*s [term frequencies] for a given word is binomial, which approximates to a Poisson under these conditions.

  The eliteness model can be seen as a simple topical model which causes variation in the unigram distributions. The author is assumed first to choose which topics to cover, i.e., which terms to treat as elite and which not. This defines specific probabilities for the unigram model, and the author then fills the word-positions according to this chosen model.

  This generative version of the 2-Poisson model (that is, a model for how documents are generated) ties it very closely with the language models.

  The model depends fairly crucially on the notion that all documents are of the same (fixed) length. [Robertson 2009]

Okapi BM25

Term Weighting

$$\omega_{\text{BM25}}(t,d,D) = \omega_{\text{dtf}}(t,d) \cdot \omega_{\text{qtf}}(t,q) \cdot \omega_{\text{BIM}}(t,D)$$

$$\omega_{\text{dtf}}(t,d) = \frac{(k_1+1)\cdot \mathit{tf}(t,d)}{k_1((1-b)+b\cdot\frac{|d|}{|d|_{\text{avg}}})+\mathit{tf}(t,d)}$$

$$\omega_{\text{qtf}}(t,q) = \frac{(k_2+1)\cdot \mathit{tf}(t,q)}{k_2+\mathit{tf}(t,q)}$$

$$\omega_{\text{BIM}}(t,D) = \log\frac{|D|-\mathit{df}(t,D)}{\mathit{df}(t,D)}$$

Okapi BM25

Term Weighting

$$\omega_{\text{BM25}}(t,d,D) = \omega_{\text{dtf}}(t,d) \cdot \omega_{\text{qtf}}(t,q) \cdot \omega_{\text{BIM}}(t,D)$$

$$\omega_{\text{dtf}}(t,d) = \frac{(k_1+1)\cdot \mathit{tf}(t,d)}{k_1((1-b)+b\cdot\frac{|d|}{|d|_{\text{avg}}})+\mathit{tf}(t,d)}$$

$$\omega_{\text{qtf}}(t,q) = \frac{(k_2+1)\cdot \mathit{tf}(t,q)}{k_2+\mathit{tf}(t,q)}$$

$$\omega_{\text{BIM}}(t,D) = \log\frac{|D|-\mathit{df}(t,D)}{\mathit{df}(t,D)}$$

Okapi BM25

Term Weighting

$$\omega_{\text{BM25}}(t,d,D) = \omega_{\text{dtf}}(t,d) \cdot \omega_{\text{qtf}}(t,q) \cdot \omega_{\text{BIM}}(t,D)$$

$$\omega_{\text{dtf}}(t,d) = \frac{(k_1+1)\cdot \mathit{tf}(t,d)}{k_1((1-b)+b\cdot\frac{|d|}{|d|_{\text{avg}}})+\mathit{tf}(t,d)}$$

$$\omega_{\text{qtf}}(t,q) = \frac{(k_2+1)\cdot \mathit{tf}(t,q)}{k_2+\mathit{tf}(t,q)}$$

$$\omega_{\text{BIM}}(t,D) = \log\frac{|D|-\mathit{df}(t,D)}{\mathit{df}(t,D)}$$

Okapi BM25

Term Weighting

$$\omega_{\text{BM25}}(t,d,D) = \omega_{\text{dtf}}(t,d) \cdot \omega_{\text{qtf}}(t,q) \cdot \omega_{\text{BIM}}(t,D)$$

$$\omega_{\text{dtf}}(t,d) = \frac{(k_1+1)\cdot \mathit{tf}(t,d)}{k_1((1-b)+b\cdot\frac{|d|}{|d|_{\text{avg}}})+\mathit{tf}(t,d)}$$

$$\omega_{\text{qtf}}(t,q) = \frac{(k_2+1)\cdot \mathit{tf}(t,q)}{k_2+\mathit{tf}(t,q)}$$

$$\omega_{\text{BIM}}(t,D) = \log\frac{|D|-\mathit{df}(t,D)}{\mathit{df}(t,D)}$$

Saturation:

- The eliteness of a term does not grow linearly with its frequency. This is represented by the cumulative distribution function of a Poisson distribution.
- Normalizing term frequency by $k_1 + \mathit{tf}(t,d)$ yields a similar function for $k_1 > 0$.
- Multiplying by (k_1+1) ensures that the weights are ≥ 1.

Okapi BM25

Term Weighting

$$\omega_{\text{BM25}}(t,d,D) = \omega_{\text{dtf}}(t,d) \cdot \omega_{\text{qtf}}(t,q) \cdot \omega_{\text{BIM}}(t,D)$$

$$\omega_{\text{dtf}}(t,d) = \frac{(k_1+1)\cdot \mathit{tf}(t,d)}{k_1((1-b)+b\cdot\frac{|d|}{|d|_{\text{avg}}})+\mathit{tf}(t,d)}$$

$$\omega_{\text{qtf}}(t,q) = \frac{(k_2+1)\cdot \mathit{tf}(t,q)}{k_2+\mathit{tf}(t,q)}$$

$$\omega_{\text{BIM}}(t,D) = \log\frac{|D|-\mathit{df}(t,D)}{\mathit{df}(t,D)}$$

Document length:

- Normalization by the average document length $|d|_{\text{avg}}$ ensures independence of datasets and implementation details.
- Why authors increase document length: verbosity or scope. The former suggests normalization, the latter not: choose $0 \leq b \leq 1$ for a mixture.

Okapi BM25

Term Weighting

$$\omega_{\text{BM25}}(t,d,D) = \omega_{\text{dtf}}(t,d) \cdot \omega_{\text{qtf}}(t,q) \cdot \omega_{\text{BIM}}(t,D)$$

$$\omega_{\text{dtf}}(t,d) = \frac{(k_1+1)\cdot \mathit{tf}(t,d)}{k_1((1-b)+b\cdot\frac{|d|}{|d|_{\text{avg}}})+\mathit{tf}(t,d)}$$

$$\omega_{\text{qtf}}(t,q) = \frac{(k_2+1)\cdot \mathit{tf}(t,q)}{k_2+\mathit{tf}(t,q)}$$

$$\omega_{\text{BIM}}(t,D) = \log\frac{|D|-\mathit{df}(t,D)}{\mathit{df}(t,D)}$$

Term frequency weighting for query:

- Like document term frequency weighting.
- Length normalization can be omitted if queries are generally short.
- Useful for query by example scenarios.

Okapi BM25

Term Weighting

$$\omega_{\text{BM25}}(t,d,D) = \omega_{\text{dtf}}(t,d) \cdot \omega_{\text{qtf}}(t,q) \cdot \omega_{\text{BIM}}(t,D)$$

$$\omega_{\text{dtf}}(t,d) = \frac{(k_1+1)\cdot \mathit{tf}(t,d)}{k_1((1-b)+b\cdot\frac{|d|}{|d|_{\text{avg}}})+\mathit{tf}(t,d)}$$

$$\omega_{\text{qtf}}(t,q) = \frac{(k_2+1)\cdot \mathit{tf}(t,q)}{k_2+\mathit{tf}(t,q)}$$

$$\omega_{\text{BIM}}(t,D) = \log\frac{|D|-\mathit{df}(t,D)}{\mathit{df}(t,D)}$$

Parameters:

- k_1 must be optimized against D; $k_1 = 1.2$ is a good value to start with.
- k_2 must be optimized against Q; in practice $0 \leq k_2 \leq 1000$, the shorter the queries, the less sensitive the overall weight is to k_2.
- b must be optimized against D; $b = 0.75$ is a good value to start with.

Okapi BM25

Term Weighting

$$\omega_{\text{BM25}}(t,d,D) \quad = \quad \omega_{\text{dtf}}(t,d) \quad \cdot \quad \omega_{\text{qtf}}(t,q) \quad \cdot \quad \omega_{\text{BIM}}(t,D)$$

$$\omega_{\text{dtf}}(t,d) \quad = \quad \frac{(k_1+1)\cdot \mathit{tf}(t,d)}{k_1((1-b)+b\cdot\frac{|d|}{|d|_{\text{avg}}})+\mathit{tf}(t,d)}$$

$$\omega_{\text{qtf}}(t,q) \quad = \quad \frac{(k_2+1)\cdot \mathit{tf}(t,q)}{k_2+\mathit{tf}(t,q)}$$

$$\omega_{\text{BIM}}(t,D) \quad = \quad \log\frac{|D|-\mathit{df}(t,D)}{\mathit{df}(t,D)}$$

Extension: BM25F (simple)

If documents have fields of varying importance, they can be weighted as follows:

$$\mathit{tf}'(t,d)=\sum_{s\in d} k_s\cdot \mathit{tf}(t,s); \qquad |d|'=\sum_{s\in d} k_s \sum_{t\in s} \mathit{tf}(t,s), \qquad |d|'_{\text{avg}}=\frac{1}{|D|}\sum_{d\in D}|d|' \ ,$$

where each s denotes a field of document d, and k_s the field-specific weight.

Okapi BM25

Term Weighting

$$\omega_{\text{BM25}}(t,d,D) \quad = \quad \omega_{\text{dtf}}(t,d) \quad \cdot \quad \omega_{\text{qtf}}(t,q) \quad \cdot \quad \omega_{\text{BIM}}(t,D)$$

$$\omega_{\text{dtf}}(t,d) \quad = \quad \frac{(k_1+1)\cdot \mathit{tf}(t,d)}{k_1((1-b)+b\cdot\frac{|d|}{|d|_{\text{avg}}})+\mathit{tf}(t,d)}$$

$$\omega_{\text{qtf}}(t,q) \quad = \quad \frac{(k_2+1)\cdot \mathit{tf}(t,q)}{k_2+\mathit{tf}(t,q)}$$

$$\omega_{\text{BIM}}(t,D) \quad = \quad \log\frac{|D|-\mathit{df}(t,D)}{\mathit{df}(t,D)}$$

Relevance Function ρ

$$\rho(\mathbf{d},\mathbf{q}) = \sum_{t\in\mathbf{q}} \omega_{\text{BM25}}(t,d,D) \ ,$$

where D is the document collection indexed.

Okapi BM25

Discussion

Advantages:

- Very good retrieval performance
- Well tunable to different retrieval scenarios
- Most terms can be precomputed at indexing time

Disadvantages:

- Departure from a rigorous theoretical probabilistic foundation