Chapter IR:V

V. Retrieval Models

- Overview of Retrieval Models
- Empirical Models
- Boolean Retrieval
- Vector Space Model
- Probabilistic Models
- Binary Independence Model
- Okapi BM25
- Hidden Variable Models
- Latent Semantic Indexing
- Explicit Semantic Analysis
- Generative Models
- Language Models
- Combining Evidence
- Web Search
- Learning to Rank

©HAGEN/POTTHAST/STEIN 2018

Probabilistic Models [Empirical Models] [Hidden Variable Models] [Generative Models]

For a pair of query q and document d, probabilistic retrieval models estimate the probability $P(\textit{relevance}(d, q) = 1 \mid \texttt{data})$ that d will be considered relevant for q.

The $\texttt{data}$ may include anything available to a retrieval system, including query q, document d, user context, user profile, etc.

The relevance function $\rho(d, q) = P(\textit{relevance}(d, q) = 1 \mid \texttt{data})$ ranks documents by their probability of being relevant.

Relevance is assumed a binary property of an individual document d given query q.

Discriminating factors of probabilistic models:

- Sample space from which probabilities are derived
- Estimation method for $P(\textit{relevance}(d, q) = 1 \mid \texttt{data})$

©HAGEN/POTTHAST/STEIN 2018

Probabilistic Models

Probability Ranking Principle [Robertson 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.

©HAGEN/POTTHAST/STEIN 2018

Probabilistic Models

Probability Ranking Principle [Robertson 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.

The truth about the relevance of a document is unknown to the retrieval system.

It can only reason about the $\texttt{data}$ at hand: the query q, the document d in question, and all supplementary data.

The probability $P(\textit{relevance}(d, q) = 1 \mid \texttt{data})$ corresponds to the proportion of all users in similar contexts querying for q, who would consider document d relevant.

It follows that ranking in decreasing order of that probability is the optimal strategy under uncertain and incomplete knowledge of the true relevance relation.

This maximizes commonly used effectiveness metrics: precision, recall, and others.

Most salient question: How to estimate $P(\textit{relevance}(d, q) = 1 \mid \texttt{data})$?

©HAGEN/POTTHAST/STEIN 2018

Remarks:

- The above rendition of the probability ranking principle is a more concise version of Cooper's original one, published by [Robertson 1977].
- The probability ranking principle is not universal since counterexamples can be formulated where ranking in order of decreasing relevance probability leads to suboptimal rankings (think of user relevance: someone has seen a document before such that it is maybe less relevant to the user but this might not be part of `data`).

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

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

Document representations $\mathbf{D}$.

The set of index terms $T = \{t_1, \ldots, t_m\}$ is composed of the word stems of the vocabulary of a document collection.

The representation $\mathbf{d}$ of a document d is a function from T to $\{0, 1\}$, where $\mathbf{d}(t_i) = 1$ is interpreted as "term t_i present in d", and $\mathbf{d}(t_i) = 0$ as "term t_i absent from d".

Query representations $\mathbf{Q}$.

A query representation $\mathbf{q}$ is constructed like a document representation.

Relevance function ρ.

The relevance function $\rho(\mathbf{d}, \mathbf{q}) = P(\mathit{relevance}(d, q) = 1 \mid \mathbf{d}, \mathbf{q})$

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

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

Document representations $\mathbf{D}$.

The set of index terms $T = \{t_1, \ldots, t_m\}$ is composed of the word stems of the vocabulary of a document collection.

The representation $\mathbf{d}$ of a document d is a function from T to $\{0, 1\}$, where $\mathbf{d}(t_i) = 1$ is interpreted as "term t_i present in d", and $\mathbf{d}(t_i) = 0$ as "term t_i absent from d".

Query representations $\mathbf{Q}$.

A query representation $\mathbf{q}$ is constructed like a document representation.

Relevance function ρ.

The relevance function $\rho(\mathbf{d}, \mathbf{q}) = P(\textit{relevance}(d, q) = 1 \mid \mathbf{d}, \mathbf{q})$

©HAGEN/POTTHAST/STEIN 2018

Remarks:

- The model is also known as Okapi model, City model, 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.

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Relevance Function ρ: Derivation

Let r denote a random variable indicating the true (binary) *relevance*(d, q), and $\mathbf{d}, \mathbf{q}$ random variables representing 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.

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Relevance Function ρ: Derivation

Let r denote a random variable indicating the true (binary) *relevance*(d, q), and $\mathbf{d}, \mathbf{q}$ random variables representing 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.

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Relevance Function ρ: Derivation

Let r denote a random variable indicating the true (binary) *relevance*(d, q), and $\mathbf{d}, \mathbf{q}$ random variables representing 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 $\mathbf{d}$.

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Relevance Function ρ: Derivation

Let $\mathbf{d}(t)$ denote a random variable indicating if t occurs in d.

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

(4) Assuming independence between terms.

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Relevance Function ρ: Derivation

Let $\mathbf{d}(t)$ denote a random variable indicating 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 $t \in T : \mathbf{d}(t) = 0$.

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Relevance Function ρ: Derivation

Let $\mathbf{d}(t)$ denote a random variable indicating 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 $s_t = P(\mathbf{d}(t)=1 \mid r=0,\mathbf{q})$.

©HAGEN/POTTHAST/STEIN 2018

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 that if $t \notin \mathbf{q}$, then $p_t = s_t$ (i.e., t is equally likely to occur in relevant and irrelevant documents). Renders all corresponding factors idempotent.

©HAGEN/POTTHAST/STEIN 2018

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 that if $t \notin \mathbf{q}$, then $p_t = s_t$ (i.e., t is equally likely to occur in relevant and irrelevant documents). Renders all corresponding factors idempotent.

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

©HAGEN/POTTHAST/STEIN 2018

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 that if $t \notin \mathbf{q}$, then $p_t = s_t$ (i.e., t is equally likely to occur in relevant and irrelevant documents). Renders all corresponding factors idempotent.

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

©HAGEN/POTTHAST/STEIN 2018

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 that if $t \notin \mathbf{q}$, then $p_t = s_t$ (i.e., t is equally likely to occur in relevant and irrelevant documents). Renders all corresponding factors idempotent.

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

©HAGEN/POTTHAST/STEIN 2018

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 that if $t \notin \mathbf{q}$, then $p_t = s_t$ (i.e., t is equally likely to occur in relevant and irrelevant documents). Renders all corresponding factors idempotent.

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

©HAGEN/POTTHAST/STEIN 2018

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 that if $t \notin \mathbf{q}$, then $p_t = s_t$ (i.e., t is equally likely to occur in relevant and irrelevant documents). Renders all corresponding factors idempotent.

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

©HAGEN/POTTHAST/STEIN 2018

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 that if $t \notin \mathbf{q}$, then $p_t = s_t$ (i.e., t is equally likely to occur in relevant and irrelevant documents). Renders all corresponding factors idempotent.

(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 $\mathbf{d}$.

©HAGEN/POTTHAST/STEIN 2018

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.

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Relevance Function ρ: Derivation

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

$$= \sum_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \underbrace{\log \frac{p_t(1-s_t)}{s_t(1-p_t)}}_{:= \omega^{\text{RSJ}}}$$

(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 irrelevant 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

©HAGEN/POTTHAST/STEIN 2018

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 $1 - p_t$.

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Relevance Function ρ: Estimation

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

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

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

(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). Since most documents in D are irrelevant, this estimation is quite accurate.

©HAGEN/POTTHAST/STEIN 2018

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]

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Relevance Function ρ: Example

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

Token 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

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

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

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

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Relevance Function ρ: Example

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

Token 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

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

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

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

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Relevance Function ρ: Example

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

Token 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	

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Relevance Function ρ: Example

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

Token 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

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

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

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

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Relevance Function ρ: Example

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

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

Token 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

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Relevance Function ρ: Example

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

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

Token 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

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

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

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Relevance Function ρ: Example

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

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

Token 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

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

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?

©HAGEN/POTTHAST/STEIN 2018

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 irrelevant documents.
5. Terms in $\mathbf{q}$ are equally likely to occur or not to occur in a relevant document d.
6. The set of irrelevant documents is represented by the entire collection.

©HAGEN/POTTHAST/STEIN 2018

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]

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Query Refinement: Relevance Feedback

Let R denote a result set for q, where R^+ and R^- are subsets of relevant and irrelevant results from relevance feedback, and R_t the subset containing term t.

$$\sum_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \log \frac{p_t(1-s_t)}{s_t(1-p_t)} \overset{\text{rank}}{=} \sum_{\substack{t \in \mathbf{q}: \\ t \in \mathbf{d}}} \log \frac{(|R_t^+| + 0.5)(|R_t^-| - |R^-| - 0.5)}{(|R_t^-| + 0.5)(|R_t^+| - |R^+| - 0.5)} \tag{13}$$

(13) Assumption that partial knowledge of the true relevance function r is at hand:

$t \in \mathbf{q}$	Relevant	Irrelevant	Σ
$t \in \mathbf{d}$	$\lvert R_t^+ \rvert$	$\lvert R_t^- \rvert$	$\lvert R_t \rvert$
$t \notin \mathbf{d}$	$\lvert R^+ - R_t^+ \rvert$	$\lvert R^- - R_t^- \rvert$	$\lvert R - R_t \rvert$
Σ	$\lvert R^+ \rvert$	$\lvert R^- \rvert$	$\lvert R \rvert$

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

$$p_t = \frac{|R_t^+| + 0.5}{|R^+| + 1.0}; \qquad s_t = \frac{|R_t^-| + 0.5}{|R^-| + 1.0}.$$

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Query Refinement: Relevance Feedback Example

$$q = (\mathtt{b}\ \mathtt{g}\ \mathtt{h})$$

Term t	a	**b**	c	d	e	f	**g**	**h**
$\lvert R_t^+ \rvert$	1	2	1	1	1	1	0	0
$\lvert R_t^- \rvert$	1	3	1	2	2	0	2	0
p_t	0.5	0.8	0.5	0.5	0.5	0.5	0.2	0.2
s_t	0.4	0.9	0.4	0.6	0.6	0.1	0.6	0.1

$$\begin{aligned} R &= R^+ \cup R^- \\ R^+ &= \{\ d_1 = (\mathtt{a}\ \mathbf{b}\ \mathtt{c}\ \mathbf{b}\ \mathtt{d}), \\ &\qquad d_2 = (\mathbf{b}\ \mathtt{e}\ \mathtt{f}\ \mathbf{b})\ \} \\ R^- &= \{\ d_3 = (\mathbf{b}\ \mathbf{g}\ \mathtt{c}\ \mathtt{d}), \\ &\qquad d_4 = (\mathbf{b}\ \mathtt{d}\ \mathtt{e}), \\ &\qquad d_5 = (\mathtt{a}\ \mathbf{b}\ \mathtt{e}\ \mathbf{g})\ \} \end{aligned}$$

$$d_6 = (\mathbf{b}\ \mathbf{g}\ \mathbf{h})$$

$$\begin{aligned} \rho(\mathbf{d}_6, \mathbf{q}) &= \underbrace{\log \frac{0.8 \cdot (1-0.9)}{0.9 \cdot (1-0.8)}}_{t=\mathbf{b}} + \underbrace{\log \frac{0.2 \cdot (1-0.6)}{0.6 \cdot (1-0.2)}}_{t=\mathbf{g}} + \underbrace{\log \frac{0.2 \cdot (1-0.1)}{0.1 \cdot (1-0.2)}}_{t=\mathbf{h}} \\ &= \log 0.44 + \log 0.17 + \log 2.25 \\ &= -0.3522 + -0.7722 + 0.3522 \\ &= -0.7722 \end{aligned}$$

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Query Refinement: Relevance Feedback

Initialization of $p_t = 0.5$ and s_t as in the absence of relevance feedback. Updating of p_t based only on a given R may be unreliable, especially when $|R|$ is small. A better update formula for p_t is as follows:

$$p'_t = \frac{|R_t^+| + \alpha p_t}{|R^+| + \alpha},$$

where α adjusts the contribution of the previous feedback cycle (e.g., $\alpha = |R|$).

©HAGEN/POTTHAST/STEIN 2018

Binary Independence Model

Query Refinement: Relevance Feedback

Initialization of $p_t = 0.5$ and s_t as in the absence of relevance feedback. Updating of p_t based only on a given R may be unreliable, especially when $|R|$ is small. A better update formula for p_t is as follows:

$$p'_t = \frac{|R_t^+| + \alpha p_t}{|R^+| + \alpha},$$

where α adjusts the contribution of the previous feedback cycle (e.g., $\alpha = |R|$).

New terms may be added to query q from relevant documents in R^+. A term weighting scheme $\omega'(t)$ ranks terms, determining those that will maximally increase the difference in average score between relevant and non-relevant documents:

$$\begin{aligned}
\omega'(t) &= (P(\mathbf{d}(t) = 1 \mid r = 1) - P(\mathbf{d}(t) = 1 \mid r = 0))\, \omega(t) && (1)\\
&\approx P(\mathbf{d}(t) = 1 \mid r = 1)\, \omega(t) && (2)\\
&\approx \frac{|R_t^+|}{|R^+|}\, \omega(t) && (3)\\
&\overset{\text{rank}}{=} |R_t^+|\, \log \frac{p_t(1 - s_t)}{s_t(1 - p_t)} && (4)
\end{aligned}$$

©HAGEN/POTTHAST/STEIN 2018

Remarks:

- Assumptions [Robertson 2009]:
  (1) Difference in average term weights between relevant and irrelevant documents.
  (2) Omission of the second probability, which is usually much smaller than the first.
  (3) Maximum likelihood estimation of the first probability as proportion of relevant documents $|R_t^+|$ containing term t among all relevant documents $|R^+|$.
  (4) Multiplication by $|R^+|$, which is a constant for a given query q's result set R.
- The top-k terms are added to $\mathbf{q}$; say, $k = 10$.

©HAGEN/POTTHAST/STEIN 2018

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

©HAGEN/POTTHAST/STEIN 2018

Okapi BM25

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

Document representations $\mathbf{D}$.

The set of index terms $T = \{t_1, \ldots, t_m\}$ is typically composed of the word stems of the vocabulary of a document collection, excluding stop words.

The representation $\mathbf{d}$ of a document d is a $|T|$-dimensional vector, where the i-th vector component of $\mathbf{d}$ corresponds to a term weight w_i of term $t_i \in T$, indicating its importance for d.

The term weighting scheme is the BM25 variant of the Best Match term weighting schemes.

Query representations $\mathbf{Q}$.

A query representation $\mathbf{q}$ is constructed like a document representation.

Relevance function ρ.

Sum of the term weights computed for the words in a query.

©HAGEN/POTTHAST/STEIN 2018

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.

©HAGEN/POTTHAST/STEIN 2018

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]

©HAGEN/POTTHAST/STEIN 2018

Okapi BM25

Term Weighting

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

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

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

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

©HAGEN/POTTHAST/STEIN 2018

Okapi BM25

Term Weighting

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

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

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

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

©HAGEN/POTTHAST/STEIN 2018

Okapi BM25

Term Weighting

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

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

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

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

©HAGEN/POTTHAST/STEIN 2018

Okapi BM25

Term Weighting

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

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

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

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

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.

©HAGEN/POTTHAST/STEIN 2018

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:

- Why authors increase document length: verbosity or scope. The former suggests normalization, the latter not; real documents are mixtures.
- Normalization by the average document length $|d|_{\text{avg}}$ ensures independence of datasets and implementation details.

©HAGEN/POTTHAST/STEIN 2018

Okapi BM25

Term Weighting

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

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

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

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

Term frequency weighting for query:

- Like document term frequency weighting, albeit without length normalization.
- Can be omitted if queries are generally short.
- Useful for query by example scenarios.

©HAGEN/POTTHAST/STEIN 2018

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.

©HAGEN/POTTHAST/STEIN 2018

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:

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

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

©HAGEN/POTTHAST/STEIN 2018

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

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

where D is the document collection indexed.

©HAGEN/POTTHAST/STEIN 2018

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 the theoretic probabilistic foundation ➜ an empirical model

©HAGEN/POTTHAST/STEIN 2018