

Probabilistic Retrieval as Divergence From Randomness

Norbert Fuhr

Content

- Basic concepts
- Models of randomness

Basic Ideas

- non-parametric models of IR
- similar to language models
- but more general framework

Basic assumptions

1. non-informative words are randomly distributed in the document set

$Prob_1$: probability of observing tf occurrences of a term in a random document

different basic probabilistic models possible

2. *elite set*: set of documents in which term occurs
regard term distribution in elite set

$Prob_2$: probability of observing tf occurrences of the term in an element of the elite set

Basic term weight

Product of two factors:

$Prob_1$ information content of the term in a document
($-\log_2 Prob_1$): corresponding amount of information.

$Prob_2$ information gain of the term wrt. its 'elite' set
The less the term is expected, the higher is the amount of information gained:
weight: $(1 - Prob_2)$

Term weight:

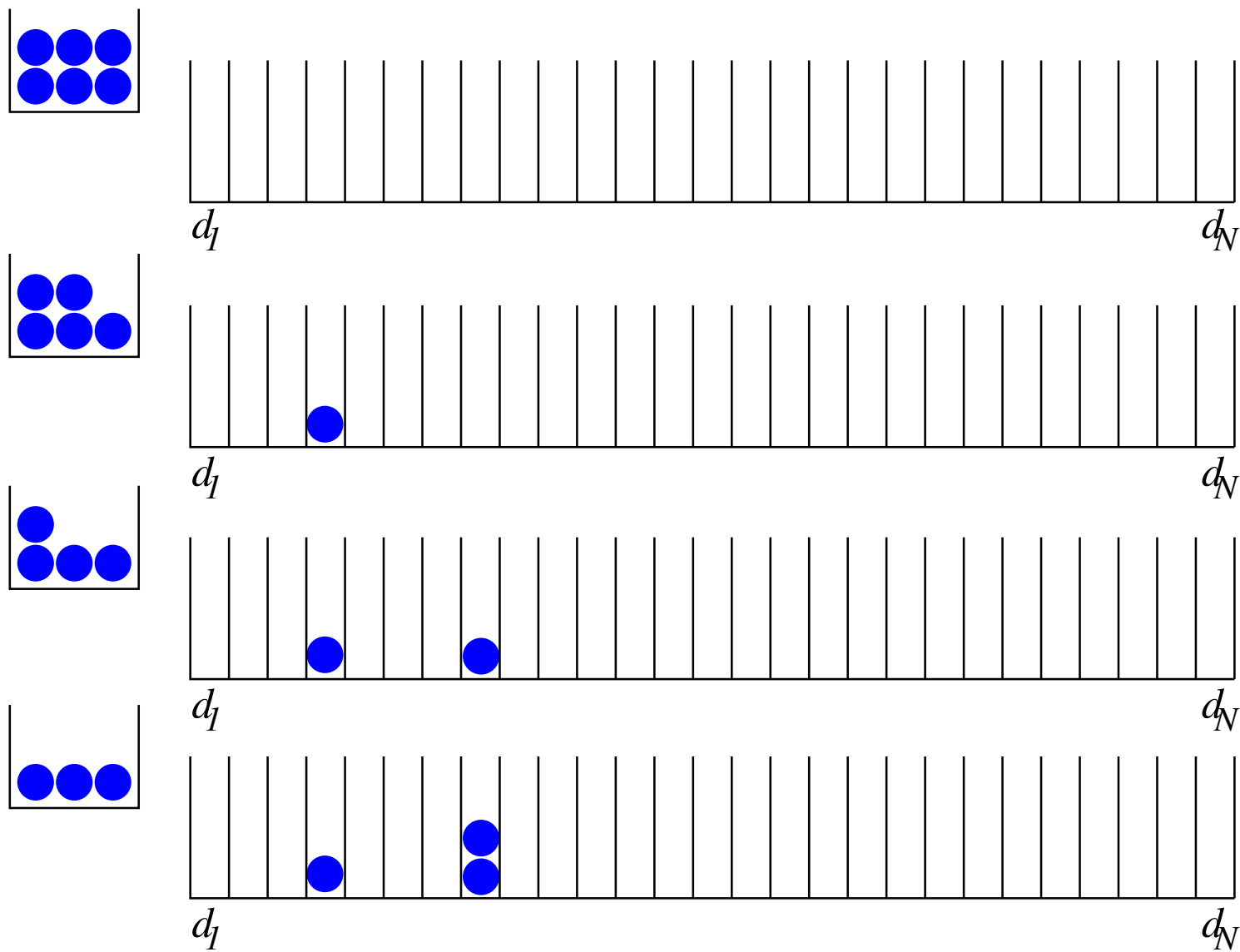
$$w = (1 - Prob_2) \cdot (-\log_2 Prob_1) = Inf_2 \cdot Inf_1$$

Models of randomness

depending on assumptions about event space
(definition of equiprobable events)

- Binomial model
- Bose-Einstein model

Binomial model



Binomial model

Basic event: occurrence of a single term in a document

Bernoulli process with $p = \frac{1}{N}$,

where $N = \#$ documents in the collection

Example: $N = 1024$ documents in the collection,

$F = 10$ total occurrences of a term (collection frequency)

$tf = 4$ occurrences of the term in a single document

Probability of randomness:

$$B(1024, 10, 4) = \binom{10}{4} p^4 q^6 = 0.000000000019$$

where $p = \frac{1}{1024}$ and $q = \frac{1023}{1024}$

Binomial model

General form:

$$Prob_1(tf) = Prob_1 = B(N, F, tf) = \binom{F}{tf} p^{tf} q^{F-tf}$$

with $p = \frac{1}{N}$ and $q = \frac{N-1}{N}$.

Binomial model does not consider size of elite set!

Approximation of binomial model: Poisson

assume that $p \rightarrow 0$ for $N \rightarrow \infty$, but $\lambda = p \cdot F = \text{constant}$
($\lambda = \text{expected } \# \text{ occurrences of the term in a document}$)
→ Poisson process:

$$B(N, F, tf) \approx \text{Poiss}(\lambda, tf) = \frac{e^{-\lambda} \lambda^{tf}}{tf!}$$

$$\begin{aligned} \text{Inf}_1(tf) &= -\log_2 B(N, F, tf) \\ &\approx -\log_2 \frac{e^{-\lambda} \lambda^{tf}}{tf!} \\ &= -tf \cdot \log_2 \lambda + \lambda \cdot \log_2 e + \log_2(tf!) \\ &\approx tf \cdot \log_2 \frac{tf}{\lambda} + \left(\lambda + \frac{1}{12tf + 1} - tf \right) \log_2 e + 0.5 \log_2(2\pi \cdot tf) \end{aligned}$$

using Stirling's formula:

$$n! = \sqrt{2\pi} \cdot n^{n+0.5} e^{-n} e^{(12 \cdot n + 1)^{-1}}$$

Approximation of binomial model: Divergence

Let $\phi = \frac{tf}{F}$ and $p = \frac{1}{N}$

divergence of ϕ from p : $D(\phi, p) = \phi \cdot \log_2 \frac{\phi}{p} + (1 - \phi) \cdot \log_2 \frac{(1-\phi)}{(1-p)}$

$$\begin{aligned} B(N, F, tf) &= \binom{F}{tf} p^{tf} q^{F-tf} \\ &\approx \frac{2^{-F \cdot D(\phi, p)}}{\sqrt{(2\pi \cdot tf(1-\phi))}} \end{aligned}$$

(using Stirling's formula)

Bose-Einstein model

place randomly F tokens of a word in N documents
event is completely described by its occupancy numbers
(each occupancy equiprobable):

tf_1, \dots, tf_N

(tf_k : term frequency in the k -th document)

Regard occupancy problem:

all N -tuples satisfying the equation

$$tf_1 + \dots + tf_N = F \quad (1)$$

number s_1 of solutions of Equation 1:

$$s_1 = \binom{N + F - 1}{F} = \frac{(N + F - 1)!}{(N - 1)!F!} \quad (2)$$

Bose-Einstein model (2)

observe k -th document with term frequency $tf \rightarrow$
random allocation of the remaining $F - tf$ tokens in other $N - 1$
documents:

$$tf_1 + \dots + tf_{k-1} + tf_{k+1} + \dots + tf_N = F - tf \quad (3)$$

number of solutions s_2 of (3):

$$s_2 = \binom{N - 1 + (F - tf) - 1}{F - tf} = \frac{(N + F - tf - 2)!}{(N - 2)!(F - tf)!} \quad (4)$$

Bose-Einstein model (3)

Probability of randomness in Bose-Einstein model: $\frac{s_2}{s_1}$

$$\begin{aligned}
 Prob_1(tf) &= \frac{\binom{N - F - tf - 2}{F - tf}}{\binom{N + F - 1}{F}} = \frac{(N + F - tf - 2)! F! (N - 1)!}{(F - tf)! (N - 2)! (N + F - 1)!} \\
 &= \frac{(F - tf + 1) \cdot \dots \cdot F \cdot (N - 1)}{(N + F - tf - 1) \cdot \dots \cdot (N + F - 1)} \\
 &= \frac{\left(\frac{F}{N} - \frac{tf-1}{N}\right) \cdot \dots \cdot \frac{F}{N} \cdot \left(1 - \frac{1}{N}\right)}{\left(1 + \frac{F}{N} - \frac{tf+1}{N}\right) \cdot \dots \cdot \left(1 + \frac{F}{N} - \frac{1}{N}\right)} \tag{5}
 \end{aligned}$$

Approximation of Bose-Einstein model

assume that $N \gg tf$: $\frac{tf-k}{N} \approx 0$ and $\frac{k+1}{N} \approx 0$ for $k = 0, \dots, tf$

$$\begin{aligned} Prob_1(tf) &= \frac{\left(\frac{F}{N} - \frac{tf-1}{N}\right) \cdot \dots \cdot \frac{F}{N} \cdot \left(1 - \frac{1}{N}\right)}{\left(1 + \frac{F}{N} - \frac{tf+1}{N}\right) \cdot \dots \cdot \left(1 + \frac{F}{N} - \frac{1}{N}\right)} \\ &\approx \frac{\frac{F}{N} \cdot \dots \cdot \frac{F}{N} \cdot 1}{\left(1 + \frac{F}{N}\right) \cdot \dots \cdot \left(1 + \frac{F}{N}\right)} = \frac{\left(\frac{F}{N}\right)^{tf}}{\left(1 + \frac{F}{N}\right)^{tf+1}} \\ &= \left(\frac{1}{1 + \frac{F}{N}}\right) \cdot \left(\frac{\frac{F}{N}}{1 + \frac{F}{N}}\right)^{tf} \end{aligned} \tag{6}$$

Approximation of Bose-Einstein model (2)

$$Prob_1(tf) = \left(\frac{1}{1 + \frac{F}{N}}\right) \cdot \left(\frac{\frac{F}{N}}{1 + \frac{F}{N}}\right)^{tf}$$

$\lambda = \frac{F}{N}$: mean term frequency in the collection

probability that a term occurs tf times in a document:

$$Prob_1(tf) \approx \left(\frac{1}{1 + \lambda}\right) \cdot \left(\frac{\lambda}{1 + \lambda}\right)^{tf} \quad (7)$$

(geometric distribution with probability $p = \frac{1}{1 + \lambda}$.)

First normalization

resizing the information content by the *aftereffect* of sampling:
if a document contains at least one occurrence of the term, the probability of observing more occurrences in this document is higher than the probability of the first occurrence

Assumption:

probability that the observed term contributes to select a *relevant* document is high, if the probability of encountering one more token of the same term in a relevant document is similarly high.

$$Prob_2(tf) = p(tf + 1 | tf, d)$$

First normalization: Laplace

$$p(tf + 1|tf, d) \approx \frac{tf + 1}{tf + 2}$$

replacing tf by $tf - 1$:

$$Prob_2(tf) = \frac{tf}{tf + 1} \quad (8)$$

With

$$w(t, d) = (1 - Prob_2) \cdot (-\log_2 Prob_1) = Inf_2 \cdot Inf_1$$

we get the L normalization:

$$w_L(t, d) = \frac{1}{tf + 1} \cdot Inf_1(tf) \quad (9)$$

First normalization: Bernoulli

- add a new token of the term to the collection: $F \rightarrow F + 1$
- compute probability that addtl. token falls into the observed document: $tf \rightarrow tf + 1$
- regard elite set only (n instead of N)
- compare $B(n, F + 1, tf + 1)$ with $B(n, F, tf)$:

$$Prob_2 = 1 - \frac{B(n, F + 1, tf + 1)}{B(n, F, tf)} = 1 - \frac{F + 1}{n \cdot (tf + 1)}$$

→ term weight with B normalization:

$$w_B(t, d) = \frac{B(n, F + 1, tf + 1)}{B(n, F, tf)} \cdot Inf_1(tf) = \frac{F + 1}{n \cdot (tf + 1)} \cdot Inf_1(tf)$$

Second normalization: document length

$l(d)$ - document length

$\rho(l)$ - term density function

Hypotheses:

H1 The distribution of a term is uniform in the document: $\rho(l) = c$

H2 The term frequency density $\rho(l)$ is a decreasing function of the length: $\rho(l) = c/l$.

where c is determined by $tf = \int_0^{l(d)} \rho(l) dl$

Length normalization

map tf onto normalised frequency tfn

(avl – average document length in the collection)

$$tfn = \int_{l(d)}^{l(d)+avl} \rho(l) dl \quad (10)$$

yields:

$$\text{H1 } tfn = tf \cdot \frac{avl}{l(d)}$$

$$\text{H2 } tfn = tf \cdot \log_2\left(1 + \frac{avl}{l(d)}\right)$$

Experimental results

- almost all variants of the model give very good results
- different approximations of the same basic model are indistinguishable
- Bose-Einstein model slightly better than Binomial model
- term frequency normalization H2 (decreasing term density) better than H1 (uniform term density)
- First normalization variants L and B give similar results