UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Ranking Models

Information-Theoretic Models

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

Probabilistic ranking

A range of models
- Probabilistic relevance models
- Language models

Key distinguishing assumptions
- How to estimate the informativeness of a term
- How to regulate the influence of document length

Probabilistic relevance models

Term informativeness
- How much observing a term-document pair contributes to observing relevance
- $P(G|t,d)$, which boils down to tf-idf

Document length
- Either ignored (BIM) or heuristic (BM25)

Language models

Term informativeness
- How much observing the language of a document contributes to observing a query term
- $P(t|\theta_d)$, which boils down to smoothed tf

Document length
- Controlled via Bayesian smoothing

An information-theoretic look

The informativeness of a term occurrence is proportional to the amount of information it carries, with random occurrences being little informative
- Specialty terms: occur non-randomly
- Non-specialty terms: occur randomly

Can we measure random-ness?

Divergence from randomness (DFR)

> *The more the divergence of the frequency of a word t in a document d compared to its frequency in the collection, the more the information carried by t in d.*
>
> ∘ Amati and van Rijsbergen, 2002

Basic assumption #1

A term that carries little information is assumed to be randomly distributed over the whole collection C

- Given a term t, its probability distribution over the whole collection is referred to as $P_1(t|C)$
- The amount of information associated with this distribution is given by $-\log P_1(t|C)$

Basic assumption #2

An informative term is frequent in its **elite set** – the set of documents where the term occurs

- Given a term t, its probability distribution in an element d of the elite set is referred to as $P_2(t|d)$
- The less the term is expected, the higher is the amount of information gained: $1 - P_2(t|d)$

DFR scoring

General scoring

- $f(q, d) = \sum_{t \in q} w_{t,q} w_{t,d}$

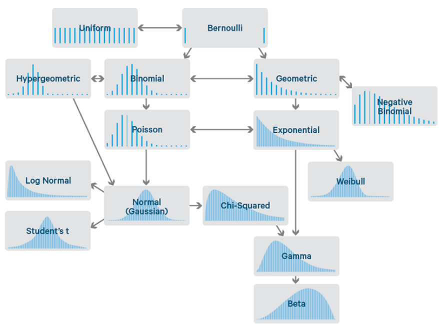
Where

- $w_{t,q} = \mathrm{tf}_{t,q} / \max_{t_i \in q} \mathrm{tf}_{t_i,q}$
- $w_{t,d} = \mathrm{inf}_1 \, \mathrm{inf}_2$
  $= -\log P_1(t|C) \times \left(1 - P_2(t|d)\right)$

Term weighting

Three steps

- Select a basic randomness model
- Apply the first normalization
- Normalize term frequencies



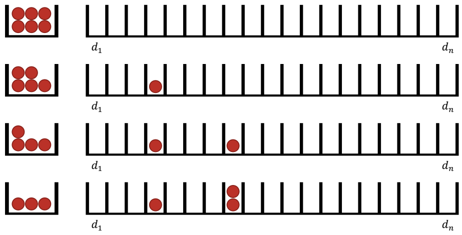
Credit: Sean Owen

Basic randomness model: $P_1(t|C)$

To compute the distribution of terms in the collection, distinct probability models can be considered

- Binomial (→ Poisson) distribution
- Bose-Einstein (→ Geometric) distribution
- Hypergeometric distribution

Binomial model (B)



Binomial model (B)

Basic event: occurrence of a single term in a document

- Bernoulli process with $p = 1/n$, for n documents

Example: $n = 1024$, $\mathrm{tf}_{t,C} = 10$, $\mathrm{tf}_{t,d} = 4$

- $P_1 = B(1024, 10, 4)$

$$= \binom{10}{4} p^4 (1-p)^6$$
$$= 0.00000000019$$

Binomial model (B)

General form

- $P_1 = B(n, \mathrm{tf}_{t,C}, \mathrm{tf}_{t,d})$

$$= \binom{\mathrm{tf}_{t,C}}{\mathrm{tf}_{t,d}} p^{\mathrm{tf}_{t,d}} (1-p)^{\mathrm{tf}_{t,C} - \mathrm{tf}_{t,d}}$$

Poisson approximation (P)

Let λ be the expected frequency of t in d

- $\lambda = \frac{\mathrm{tf}_{t,C}}{n}$ (constant)

For $n \to \infty$ ($p = 1/n \to 0$)

- $B(n, \mathrm{tf}_{t,C}, \mathrm{tf}_{t,d}) \approx Poiss(\lambda, \mathrm{tf}_{t,d})$

$$= \frac{e^{-\lambda} \lambda^{\mathrm{tf}_{t,d}}}{\mathrm{tf}_{t,d}!}$$

Poisson approximation (P)

$$\begin{aligned} \mathrm{inf}_1 &= -\log B(n, \mathrm{tf}_{t,C}, \mathrm{tf}_{t,d}) \\ &\approx -\log Poiss(\lambda, \mathrm{tf}_{t,d}) \\ &= -\log \frac{e^{-\lambda} \lambda^{\mathrm{tf}_{t,d}}}{\mathrm{tf}_{t,d}!} \\ &= -\mathrm{tf}_{t,d} \log \lambda + \lambda \log e + \log (\mathrm{tf}_{t,d}!) \end{aligned}$$

Poisson approximation (P)

Using Stirling's formula

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

$$\begin{aligned}\mathrm{inf}_1 \approx\ & \mathrm{tf}_{t,d} \log \frac{\mathrm{tf}_{t,d}}{\lambda} \\ & + \left(\lambda + \frac{1}{12\ \mathrm{tf}_{t,d} + 1} - \mathrm{tf}_{t,d}\right) \log e \\ & + 0.5 \log (2\pi\ \mathrm{tf}_{t,d})\end{aligned}$$

Bose-Einstein model (B_E)

Describes the number of particles with a certain energy

- In our setting, describes the probability that a document d contains $\mathrm{tf}_{t,d}$ occurrences of term t

Geometric model (G)

B-E can be approximated by a geometric distribution

- $P(t|C) = p(1-p)^{\mathrm{tf}_{t,d}}$, where where $p = 1/(1+\lambda)$

The amount of information associated with term t in the collection can then be computed as

$$\mathrm{inf}_1 \approx -\log\left(\frac{1}{1+\lambda}\right) - \mathrm{tf}_{t,d} \log\left(\frac{\lambda}{1+\lambda}\right)$$

First normalization: $P_2(t|d)$

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

Laplace's law of succession (L)

Useful when we have no advance knowledge of how many tokens of a term should occur in a relevant document of arbitrary large size

$$\begin{aligned}P_2 &= P(\mathrm{tf}_{t,d} + 1 | \mathrm{tf}_{t,d}, d) \\ &\approx \frac{\mathrm{tf}_{t,d} + 1}{\mathrm{tf}_{t,d} + 2} \approx \frac{\mathrm{tf}_{t,d}}{\mathrm{tf}_{t,d} + 1} \quad \text{replacing } \mathrm{tf}_{t,d} \text{ by } \mathrm{tf}_{t,d} - 1\end{aligned}$$

Laplace's law of succession (L)

$$\begin{aligned}\mathrm{inf}_2 &= 1 - P_2(t|d) \\ &\approx 1 - \frac{\mathrm{tf}_{t,d}}{\mathrm{tf}_{t,d} + 1} \\ &= \frac{1}{\mathrm{tf}_{t,d} + 1} \quad \text{tf saturation effect}\end{aligned}$$

First normalization: Bernoulli (B)

Add a new token to the collection: $\text{tf}_{t,C} \rightarrow \text{tf}_{t,C} + 1$

- Compute probability that additional token falls into the observed documents: $\text{tf}_{t,d} \rightarrow \text{tf}_{t,d} + 1$

Compare $B\big(n_t, \text{tf}_{t,C} + 1, \text{tf}_{t,d} + 1\big)$ vs $B\big(n_t, \text{tf}_{t,C}, \text{tf}_{t,d}\big)$ on the elite set only (n_t instead of n)

First normalization: Bernoulli (B)

$$P_2 = \frac{B\big(n_t, \text{tf}_{t,C} + 1, \text{tf}_{t,d} + 1\big)}{B\big(n_t, \text{tf}_{t,C}, \text{tf}_{t,d}\big)}$$

$$= \frac{\text{tf}_{t,C} + 1}{n_t(\text{tf}_{t,d} + 1)}$$

$$\text{inf}_2 = 1 - \frac{\text{tf}_{t,C} + 1}{n_t(\text{tf}_{t,d} + 1)}$$

tf saturation effect

idf effect

Second normalization: document length

Formulations thus far do not take into account the length of document d

- Solution: normalize term frequency $\text{tf}_{t,d}$

H1. $\text{tfn}_{t,d} = \text{tf}_{t,d} \frac{avl}{l_d}$

H2. $\text{tfn}_{t,d} = \text{tf}_{t,d} \log\left(1 + \gamma \frac{avl}{l_d}\right)$

Example model

PL2 [Amati, 2003]

- Randomness model: Poisson
- First normalization: Laplace
- Second normalization: H2

Many other effective models

Hypergeometric model (H)

Binomial distribution describes the probability of observing $\text{tf}_{t,d}$ after $\text{tf}_{t,C}$ independent draws

- $\text{tf}_{t,C}$ tokens are sampled ***with*** replacement

Hypergeometric distribution describes the probability of observing $\text{tf}_{t,d}$ after $\text{tf}_{t,C}$ non-independent draws

- $\text{tf}_{t,C}$ tokens are sampled ***without*** replacement

Hypergeometric model (H)

Because draws are not independent, the probability of observing a further token in a document is reduced

- In practice, no need for length normalization
- Also, no hyperparameter tuning

DPH [Amati et al. 2007]

- Very effective in web search tasks

Summary

Almost all variants of the model give very good results
- Poisson model slightly better than Binomial
- First normalization variants L and B give similar results
- Term frequency normalization H2 better than H1
- Hypergeometric model effective and parameter free

Divergence from independence (DFI)

Independence rather than randomness
- Non-specialty terms: occur at a more or less constant rate relative to other terms across documents

Independence quantification is distribution-free
- Non-parametric counterpart of DFR models
- Also very effective in practice

References

Modern Information Retrieval, Sec. 3.5.3
Baeza-Yates and Ribeiro-Neto

Common Probability Distributions: The Data Scientist's Crib Sheet
Sean Owen, Cloudera Engineering Blog 2017

References

Probabilistic models of information retrieval based on measuring the divergence from randomness
Amati and van Rijsbergen, ACM TOIS 2002

A nonparametric term weighting method for information retrieval based on measuring the divergence from independence
Kocabas, Dinçer, and Karaoglan, Inf. Retr. 2013

UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Coming next…

Feedback Models

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br