UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Ranking Models

Probabilistic Relevance Models

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

The ranking problem

Given
- Some evidence of the user's need

Produce
- A list of matching information items
- In decreasing order of relevance

The ranking problem

Given
- Some ~~evidence of the user's need~~ ***query***

Produce
- A list of matching ~~information items~~ ***documents***
- In decreasing order of relevance

The ranking problem

***f*(*q*, *d*)**

Ranking under uncertainty

Queries are uncertain
- Short and underspecified

Documents are uncertain
- Quality varies, natural language is ambiguous

Relevance estimates are bound to be uncertain
- Probability theory allows reasoning under uncertainty

Probabilistic ranking at a glance

Probabilistic relevance models
- Binary Independence Model (BIM)
- Best Match models (e.g., BM25)

Bayesian network models

Language models

Information-theoretic models

Probability Ranking Principle (PRP)

> *Ranking documents by decreasing probability of relevance results in optimal effectiveness, provided that probabilities are estimated (1) with certainty and (2) independently.*
>
> ◦ Robertson, 1977

Probabilistic ranking

Assume binary notion of relevance
- $R = 1$ (or R): if document *d* is relevant w.r.t query *q*
- $R = 0$ (or $\bar{R}$): otherwise

Probabilistic ranking orders documents decreasingly by their estimated probability of relevance w.r.t. query
- $f(q, d) = P(R|d, q)$

Probability recap #1: Bayes' rule

$$P(A|B) = \frac{P(B|A)\, P(A)}{P(B)}$$

A way of updating probabilities
- From the prior $P(A)$, derive the posterior $P(A|B)$
- Given evidence B and likelihood $P(B|A)$

Direct estimation

$$f(q, d) = P(R|d, q) = \frac{P(d|R, q)P(R|q)}{P(d|q)}$$

For a given query q
- How likely is d among R? $P(d|R, q)$
- How likely is R in general? $P(R|q)$
- How likely is d (regardless of R)? $P(d|q)$

Do we need to estimate all these?

Probability recap #2: odds of an event

$$O(A) = \frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1 - P(A)}$$

Provide a ”multiplier” for how probabilities change
- Probability of an event occurring vs.
  Probability of the event not occurring

Order preserving transformation

Odds estimation

$$f(q,d) = O(R|d,q) = \frac{P(R|d,q)}{P(\bar{R}|d,q)}$$
$$= \frac{P(d|R,q)P(R|q)/P(d|q)}{P(d|\bar{R},q)P(\bar{R}|q)/P(d|q)}$$
$$= \frac{P(d|R,q)}{P(d|\bar{R},q)} \frac{P(R|q)}{P(\bar{R}|q)}$$

Odds estimation

$$f(q,d) = O(R|d,q) \propto \frac{P(d|R,q)}{P(d|\bar{R},q)}$$

How to estimate $P(d|R,q)$?

Estimating $P(d|R,q)$

To make a probabilistic ranking strategy precise, we need to estimate the contribution of individual terms

- Find measurable term-level statistics that affect judgments about document relevance
- Combine these statistics to estimate the probability of relevance of the entire document

Binary independence assumption

"Binary"

- Documents and queries as incidence vectors
  - $q = (x_1, x_2, \dots, x_{|V|})$ with $x_i \in \{0,1\}$
  - $d = (y_1, y_2, \dots, y_{|V|})$ with $y_i \in \{0,1\}$

"Independence"

- No association between terms (not true, but works)

Binary Independence Model (BIM)

$$f(q,d) = O(R|d,q) \propto \frac{P(d|R,q)}{P(d|\bar{R},q)}$$
$$\approx \prod_t \frac{P(y_t|R,q)}{P(y_t|\bar{R},q)}$$

term independence

Further simplifying assumption

- Out-of-query terms do not matter

Binary Independence Model (BIM)

$$f(q,d) = O(R|d,q) \propto \frac{P(d|R,q)}{P(d|\bar{R},q)}$$
$$\approx \prod_t \frac{P(y_t|R,q)}{P(y_t|\bar{R},q)}$$
$$\approx \prod_t \frac{P(y_t|R)}{P(y_t|\bar{R})}$$

out-of-query terms dropped

Binary Independence Model (BIM)

$$f(q,d) \approx \prod_{t\in d} \frac{P(y_t|R)}{P(y_t|\bar{R})} \quad \frac{p_t}{u_t}$$
$$\times \prod_{t\notin d} \frac{P(\bar{y}_t|R)}{P(\bar{y}_t|\bar{R})} \quad \frac{1-p_t}{1-u_t}$$

Binary Independence Model (BIM)

$$f(q,d) \approx \prod_{t\in d} \frac{p_t}{u_t} \times \prod_{t\notin d} \frac{1-p_t}{1-u_t}$$
$$\times \left(\prod_{t\in d} \frac{1-u_t}{1-p_t} \times \prod_{t\in d} \frac{1-p_t}{1-u_t} \right)$$

trick: multiply by "1"

Binary Independence Model (BIM)

$$f(q,d) \approx \prod_{t\in d} \frac{p_t}{u_t} \times \prod_{t\in d} \frac{1-u_t}{1-p_t}$$
$$\times \prod_{t\in d} \frac{1-p_t}{1-u_t} \times \prod_{t\notin d} \frac{1-p_t}{1-u_t}$$

rearranging all products

Binary Independence Model (BIM)

$$f(q,d) \approx \prod_{t\in d} \frac{p_t(1-u_t)}{u_t(1-p_t)}$$
$$\times \prod_t \frac{1-p_t}{1-u_t}$$

document-independent component dropped

Binary Independence Model (BIM)

$$f(q,d) \approx \prod_{t\in d} \frac{p_t(1-u_t)}{u_t(1-p_t)}$$
$$\propto \log \prod_{t\in d} \frac{p_t(1-u_t)}{u_t(1-p_t)}$$
$$= \sum_{t\in d} \log \frac{p_t(1-u_t)}{u_t(1-p_t)}$$

order preserving transformation

Nice, but how do we get p_t and u_t?

Probability estimates in practice

$p_t = P(y_t|R)$

- Cannot estimate without actual relevance data
  → assume constant (e.g., $p_t = 0.5$)

$u_t = P(y_t|\bar{R})$

- Non-relevant documents are the vast majority
  → assume collection statistics ($u_t = n_t/n$)

Probability estimates in practice

$$f(q,d) \approx \sum_{t \in d} \log \frac{p_t(1-u_t)}{u_t(1-p_t)}$$
$$= \sum_{t \in d} \log \frac{0.5(1-n_t/n)}{n_t/n(1-0.5)}$$
$$= \sum_{t \in d} \log \frac{n-n_t}{n_t} \approx \sum_{t \in d} \log \frac{n}{n_t}$$

How effective is BIM?

How effective is BIM?

BIM was originally designed for short catalog records of fairly consistent length (e.g., titles or abstracts)

- Works reasonably well in these contexts

For modern full-text search collections, a model should pay attention to term frequency and document length

- BM25 is sensitive to these quantities

Okapi BM25

Binary Independence Model

- $f(q,d) \approx \sum_{t \in q \cap d} \log \frac{n}{n_t}$

Okapi/BM25 [Robertson and Walker, 1994]

- $f(q,d) = \sum_{t \in q} c(t,q) \frac{(k_1+1)\, c(t,d)}{c(t,d)+k_1\left((1-b)+b\frac{|d|}{avdl}\right)} \log \frac{n+1}{n_t}$

Summary

Relevance cannot be predicted with certainty
- Probabilistic theory provides a solid framework

BIM makes a few assumptions
- Boolean representations of docs/queries/relevance
- Terms are independent of one another
- Out-of-query terms do not affect ranking

References

Introduction to Information Retrieval, Ch. 11
Manning et al., 2008

Search Engines: Information Retrieval in Practice, Ch. 7
Croft et al., 2009

The Probabilistic Relevance Framework

Robertson and Zaragoza, FnTIR 2009

UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Coming next…

Language Models

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