Lecture 10: Classical Probabilistic IR: Binary independence model

William Webber (william@williamwebber.com)

COMP90042, 2014, Semester 1, Lecture 10

What we'll learn in this lecture

Binary probabilistic models for IR

- $P(R|d, q)$
- Binary independence model

Probabilistic vs. geometric models

Fundamental calculations:

Geometric	How similar $\text{sim}(d, q)$ is document d to query q?
Probabilistic	What probability $P(R = 1 \mid d, q)$ that d is relevant to q?

Probabilistic models

Probabilistic models:

- Clearer theoretical basis that geometric
  - Particularly when considering extensions, modifications (think of "pivoted DLN")
- Very early theory (from 1970s)
- But only in 1990s did effective retrieval models develop
- Now, many probabilistic models
- This and next lecture, look at "classical" development up to BM25
- Later, language models

Bayes theorem

Bayes' theorem states:

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

E.G. M = have malaria; T = positive test; $P(T|M) = 0.8$; $P(T|\bar{M}) = 0.01$; $P(M) = 0.001$; what is $P(M|T)$?

$$\begin{aligned} P(M|T) &= \frac{0.8 \cdot 0.001}{0.8 \cdot 0.001 + 0.01 \cdot 0.999} \\ &= \frac{0.0008}{0.0008 + 0.00999} = 0.074 \end{aligned}$$

Bayes theorem

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

- $P(A)$ is the *prior* probability (distribution) of A
- We then observe evidence B
- $P(B|A)/P(B)$ is support that B provides for A
- $P(A|B)$ is *posterior* probability of A

Bayes theorem for relevance

$$P(R|d,q) = \frac{P(d|R,q)}{P(d|q)} \cdot P(R|q) \tag{1}$$

- $P(R|q)$ can be understood as proportion of documents in collection that are relevant to query
- $P(d|R,q)$ is probability that a (retrieved) document relevant to q looks like d
- $P(d|q) = P(d|R,q)\,P(R|q) + P(d|\bar{R},q)\,P(\bar{R},q)$ is probability of observing (retrieved) document, regardless of relevance

OK, but how do we go about estimating these values?

Rank-equivalence given query

Probability Ranking Principle (PRP)

- Assume output is ranking
- Further assume that relevance of documents is independence
- Then optimal ranking is by decreasing probability of relevance

- For ranking, we only care about
  - Relative probability
  - for given query
- This allows various simplifications Equation 1
- Provided they are monotonic
- i.e., for transformation $f()$,

$$P(A) > P(B) \Rightarrow f(P(A)) > f(P(B))$$

Odds-based matching score

Take odds ratio between relevance and irrelevance:

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

$\frac{P(R|q)}{P(\bar{R}|q)}$ constant given query, so can ignore:

$$\tilde{O}(R|d,q) = \frac{P(d|R,q)}{P(d|\bar{R},q)} \tag{2}$$

We have removed 2 of the 3 terms from Equation 1

Binary independence model

- How to estimate $P(d|R, q)$ and $P(d|\bar{R}, q)$?
- Must be based on attributes of d and q

Binary indendence model

Binary Doc attributes are presence of terms (not frequency)

Independence Term appearances independent given relevance

Represent:

- Document as binary vector $\vec{d}$
- Query as binary vector $\vec{q}$

BIM odds ratio

Under BIM, Equation (2) resolves to:

$$\tilde{O}(R|\vec{d},\vec{q}) = \prod_{t=1}^{|T|} \frac{P(d_t|R,\vec{q})}{P(d_t|\bar{R},\vec{q})} = \prod_{t:d_t} \frac{P(d_t|R,\vec{q})}{P(d_t|\bar{R},\vec{q})} \cdot \prod_{t:\bar{d}_t} \frac{P(\bar{d}_t|R,\vec{q})}{P(\bar{d}_t|\bar{R},q)} \quad (3)$$

Note similarity to Naive Bayes (if you know Naive Bayes).

Query terms only

Write:

$$\begin{aligned} p_t &= P(d_t|R,q) \\ u_t &= P(d_t|\bar{R},q) \end{aligned}$$

Assume $p_t = u_t$ when $q_t = 0$ (non-query terms equally likely in relevant as irrelevant documents). Then Equation 3 becomes:

$$\begin{aligned} \tilde{O}(R|\vec{d},\vec{q}) &= \prod_{t:d_t \wedge q_t} \frac{p_t}{u_t} \cdot \prod_{t:\bar{d}_t \wedge q_t} \frac{(1-p_t)}{(1-u_t)} \\ &= \prod_{t:d_t \wedge q_t} \frac{p_t(1-u_t)}{u_t(1-p_t)} \cdot \prod_{t:q_t} \frac{(1-p_t)}{(1-u_t)} \qquad (4) \end{aligned}$$

Query-doc matches only

Term $\prod_{t:q_t} \frac{(1-p_t)}{(1-u_t)}$ in Equation (4) fixed for query, can be dropped

$$\tilde{O}(R|\vec{d},\vec{q}) = \prod_{t:d_t \wedge q_t} \frac{p_t(1-u_t)}{u_t(1-p_t)} \tag{5}$$

Log transformation monotonic, changes products to sums, gives *log odds*, which we take as matching score M:

$$\begin{aligned} M(d,q) &= \log \tilde{O}(R|\vec{d},\vec{q}) = \log \prod_{t:d_t \wedge q_t} \frac{p_t(1-u_t)}{u_t(1-p_t)} && (6) \\ &= \sum_{t:d_t \wedge q_t} \log \frac{p_t}{1-p_t} + \log \frac{1-u_t}{u_t} && (7) \end{aligned}$$

Note that only terms occurring in both query and document contribute to matching score. Weight of term t is:

$$w_t = \log \frac{p_t}{1-p_t} + \log \frac{1-u_t}{u_t} \tag{8}$$

Assessement-time estimation

$$w_t = \log \frac{p_t}{1 - p_t} + \log \frac{1 - u_t}{u_t} \tag{9}$$

- Equation for w_t still depends upon random distribution functions $p_t = P(d_t|R, q)$ and $u_t = P(d_t|\bar{R}, q)$.
- Given assessed collection, p_t and u_t directly estimatable as Bernoulli ("coin-flip") distributions:

$$\begin{aligned} \hat{p}_t &= 1/|\mathcal{R}| \sum_{d \in \mathcal{R}} d_t \\ \hat{u}_t &= 1/|\mathcal{R}'| \sum_{d \in \mathcal{R}'} d_t \end{aligned}$$

But $\mathcal{R}$ (of course) unknown at retreival time. How to estimate?

Retrieval-time estimation: u_t

- Assume relevant documents rare
- Then collection statistics estimate u_t:

$$\log \frac{1 - u_t}{u_t} \approx \log \frac{N - f_t}{f_t} \approx \log \frac{N}{f_t} \qquad (10)$$

- Look familiar?

Retrieval-time estimation: p_t

- Setting p_t to 0.5 removes $p_t/(1-p_t)$
  - Relevance score of doc is just sum of IDFs
  - Plausible for binary model
- Empirical analysis[1] suggests more accurate is:

$$p_t = \frac{1}{3} + \frac{2}{3}\frac{f_t}{N} \tag{11}$$

[1] Greiff, "A theory of term weighting", *SIGIR*, 1998

Looking back and forward

Back

- Probabilistic IR models estimate $P(R|d,q)$ (or monotonic function thereof)
- Probability derived from attributes (term occurrences) of documents
- Binary independence model assumes:
  - Binary attributes (term occurs or doesn't)
  - Term occurrences independent

Looking back and forward

Forward

- Want to include term frequencies
- Two-Poisson model (next lecture) does this, leading to BM25 metric
- Language models (later in course) an alternative probabilistic IR framework

Further reading

- Chapter 11, "Probabilistic information retrieval"[2], of Manning, Raghavan, and Schutze, *Introduction to Information Retrieval*, CUP, 2009.
- Sparck Jones, Walker, and Robertson, "A Probabilistic MOdel of Information Retrieval", *IPM*, 2000.

[2] http://nlp.stanford.edu/IR-book/pdf/11prob.pdf