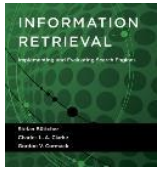
Retrieval Models

Probability Ranking Principle

Web Search

Slides based on the books:



Retrieval models

- Geometric/linear spaces
  - Vector space model
- Probability ranking principle
- Language models approach to IR
  - An important emphasis in recent work
- Probabilistic retrieval model
  - Binary independence model
  - Okapi's BM25

Recall a few probability basics

- For events A and B, the Bayes' Rule is:

$$p(A,B) = p(A|B)p(B) = p(B|A)p(A)$$

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

- Interpretation:

$$posterior = \frac{prior \cdot likelihood}{evidence} \Leftrightarrow p(A|B) = \frac{p(A)p(B|A)}{p(B)}$$

Recall a few probability basics

- Independence assumption:

$$p(A|B) = \frac{p(A)p(B|A)}{p(B)} = \frac{p(A)\prod_i p(b_i|A)}{\prod_i p(b_i)}$$

- Odds: $O(A) = \frac{p(A)}{p(\bar{A})} = \frac{p(A)}{1-p(A)}$

$$O(A|B) = \frac{p(A|B)}{p(\bar{A}|B)} = \frac{\frac{p(A)p(B|A)}{p(B)}}{\frac{p(\bar{A})p(B|\bar{A})}{p(B)}} = \frac{p(A)p(B|A)}{p(\bar{A})p(B|\bar{A})}$$

Recall a few probability basics

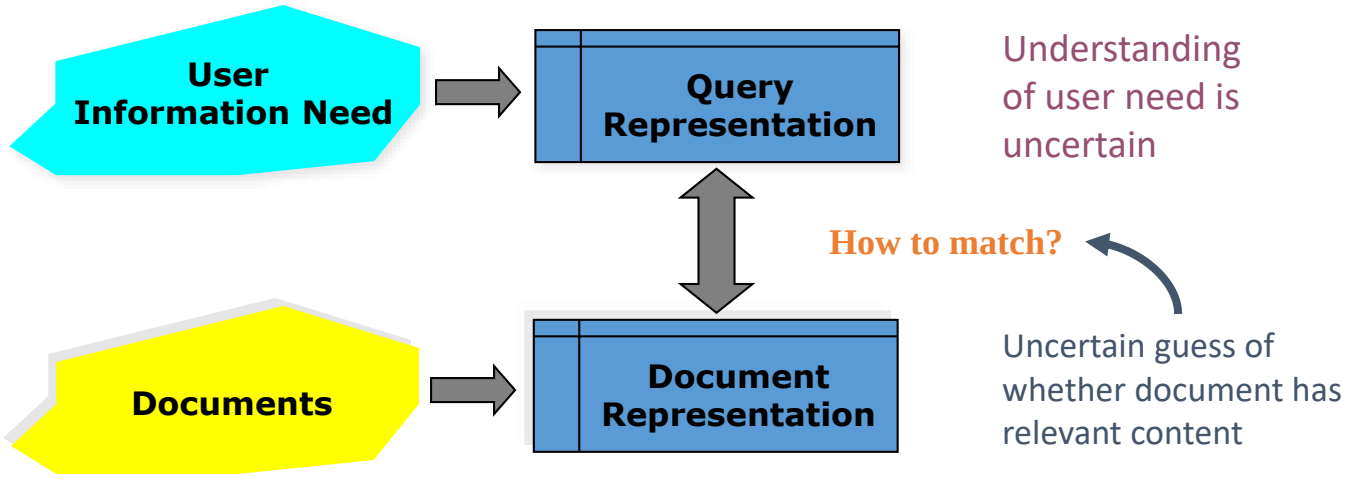
$$p(A|data) = \frac{p(A)p(data|A)}{p(data)}$$

$$p(SLB = campeão|data) = \frac{p(SLB = campeão)p(data|SLB = campeão)}{p(data)}$$

$$aposteriori = \frac{apriori \cdot verosimilhança}{evidencia}$$

Why probabilities in IR?

- In traditional IR systems, matching between each document and query is attempted in a semantically imprecise space of index terms.



Probabilities provide a principled foundation for uncertain reasoning.

Can we use probabilities to quantify our uncertainties?

The document ranking problem

- We have a collection of documents
- User issues a query
- A list of documents needs to be returned
- Ranking method is the core of an IR system:
  - In what order do we present documents to the user?
  - We want the “best” document to be first, second best second, etc….

Idea: Rank by probability of relevance of the document w.r.t. information need

Modeling relevance

P(R=1|document, query)

- Let **d** represent a document in the collection.
- Let **R** represent relevance of a document w.r.t. to a query **q**
- Let **R=1** represent relevant and **R=0** not relevant.

- Our goal is to estimate: $$p(r=1|q,d) = \frac{p(d,q|r=1)p(r=1)}{p(d,q)}$$

$$p(r=0|q,d) = \frac{p(d,q|r=0)p(r=0)}{p(d,q)}$$

Probability Ranking Principle (PRP)

- PRP in action: Rank all documents by $p(r = 1|q, d)$
  - Theorem: Using the PRP is optimal, in that it minimizes the loss (Bayes risk) under 1/0 loss
  - Provable if all probabilities correct, etc. [e.g., Ripley 1996]

$$p(r|q, d) = \frac{p(d, q|r)p(r)}{p(d, q)}$$

- Using odds, we reach a more convenient formulation of ranking :

$$\mathrm{O}(R|q, d) = \frac{p(r = 1|q, d)}{p(r = 0|q, d)}$$

Probabilistic retrieval models interpretation

- PRP in action: Rank all documents by $p(r = 1|q, d)$
  - Theorem: Using the PRP is optimal, in that it minimizes the loss (Bayes risk) under 1/0 loss
  - Provable if all probabilities correct, etc. [e.g., Ripley 1996]

$$p(r|q, d) = \frac{p(d, q|r)p(r)}{p(d, q)}$$

- Using odds, we reach a more convenient ranking formulation:

$$\mathrm{O}(R|q, d) = \frac{p(r = 1|q, d)}{p(r = 0|q, d)} \propto \frac{p(d|q, r = 1)}{p(d|q, r = 0)}$$

Language models interpretation

- In language models, we do a different formulation towards the query posterior given the document as a model.

$$\mathrm{O}(R|q,d) = \frac{p(r=1|q,d)}{p(r=0|q,d)} \propto \log \frac{p(q|d,r\,)p(r|d)}{p(q|d,\bar{r})p(\bar{r}|d)}$$

The two families of Retrieval Models

Probability Ranking Principle

$$\mathrm{O}(R|q,d) = \frac{p(r=1|q,d)}{p(r=0|q,d)}$$

Probabilistic Retrieval Models

$$O(R|q,d) \propto \frac{p(d|q,r=1)}{p(d|q,r=0)}$$

- Vector Space Model
- Binary Independent Model
- BM25

Language Models

$$O(R|q,d) \propto \log \frac{p(q|d,r\,)p(r|d)}{p(q|d,\bar{r})p(\bar{r}|d)}$$

- LM Dirichlet
- LM Jelineck-Mercer