CS47300: Web Information Search and Management

Probabilistic Retrieval Models
Prof. Chris Clifton
11 September 2020
Material adapted from course created by Dr. Luo Si, now leading Alibaba research group

Okapi BM25

- Problem: BIM favors long documents
  - More likely to contain matching terms
- Solution:
  - Inter-document frequency for “relevance”
  - Scale by document length
  - Scale by query length/term frequency
Okapi BM25

- BM25 metric, used in Okapi IR system
  - $V$=relevant documents, $VNR$=not relevant

  \[
  RSV_d = \sum_{t \in q} \left[ \frac{|V_R| + \frac{1}{2}}{|VNR_R| + \frac{1}{2}} \times \frac{(k_1 + 1)tf_{td}}{k_1 (1-b) + b \left( \frac{L_d}{L_{ave}} \right) + tf_{td}} \times \frac{(k_3 + 1)tf_{tq}}{k_3 + tf_{tq}} \right]
  \]

  - Probabilistic interpretation of IDF
  - Term frequency with normalization by length
  - Query term weighting

PRP and BIM

- Getting reasonable approximations of probabilities is possible.

- Requires restrictive assumptions:
  - **Term independence**
  - **Terms not in query don’t affect the outcome**
  - **Boolean representation of documents/queries/relevance**
  - **Document relevance values are independent**

- Some of these assumptions can be removed
- Problem: either require partial relevance information or only can derive somewhat inferior term weights
Removing term independence

- In general, index terms aren’t independent
- Dependencies can be complex
- van Rijsbergen (1979) proposed model of simple tree dependencies
  - Exactly Friedman and Goldszmidt's Tree Augmented Naive Bayes (AAAI 13, 1996)
- Each term dependent on one other
- In 1970s, estimation problems held back success of this model

And to put some faces to the names you’ve been seeing…

Karen Spärck Jones  Stephen Robertson  Keith van Rijsbergen


[Adds very little material that isn’t in van Rijsbergen or Fuhr]