PURDUE UNIVERSITY | Department of Computer Science

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

PURDUE UNIVERSITY
Department of Computer Science

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

PURDUE UNIVERSITY
Department of Computer Science

Okapi BM25

- BM25 metric, used in Okapi IR system
  - V=relevant documents, VNR=not relevant
- $RSV_d = \sum_{t \in q} \left[\left[\frac{\frac{|VR_t|+\frac{1}{2}}{|VNR_t|+\frac{1}{2}}}{\frac{df_t-|VR_t|+\frac{1}{2}}{N-df_t-|VR|+|VR_t|+\frac{1}{2}}} \right] \times \frac{(k_1+1)tf_{td}}{k_1\left((1-b)+b\left(\frac{L_d}{Lave}\right)+tf_{td}\right)} \times \frac{(k_3+1)tf_{tq}}{k_3+tf_{tq}} \right]$

Probabilistic interpretation of IDF

Term frequency with normalization by length

Query term weighting

59

PURDUE UNIVERSITY
Department of Computer Science

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

60

PURDUE UNIVERSITY
Department of Computer Science

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

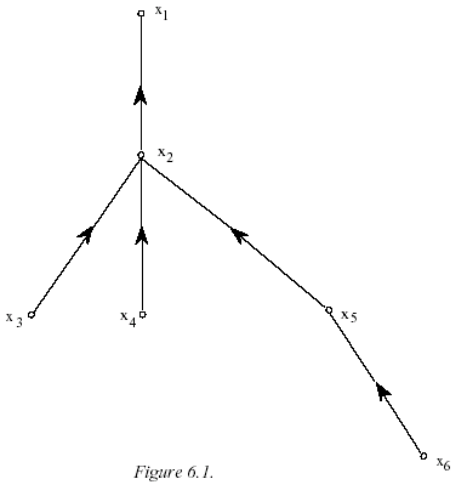

Figure 6.1.

61

PURDUE UNIVERSITY
Department of Computer Science

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


Karen Spärck Jones


Stephen Robertson


Keith van Rijsbergen

63

PURDUE UNIVERSITY
Department of Computer Science

Resources

S. E. Robertson and K. Spärck Jones. 1976. Relevance Weighting of Search Terms. *Journal of the American Society for Information Sciences* 27(3): 129–146.

C. J. van Rijsbergen. 1979. *Information Retrieval.* 2nd ed. London: Butterworths, chapter 6. [Most details of math] http://www.dcs.gla.ac.uk/Keith/Preface.html

N. Fuhr. 1992. Probabilistic Models in Information Retrieval. *The Computer Journal*, 35(3),243–255. [Easiest read, with BNs]

F. Crestani, M. Lalmas, C. J. van Rijsbergen, and I. Campbell. 1998. Is This Document Relevant? ... Probably: A Survey of Probabilistic Models in Information Retrieval. *ACM Computing Surveys* 30(4): 528–552.

http://www.acm.org/pubs/citations/journals/surveys/1998-30-4/p528-crestani/

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