11-642:
Search Engines

Document Priors

Jamie Callan
Carnegie Mellon University
callan@cs.cmu.edu

Introduction

- **Until now, the discussion of retrieval models treated a document as a bag of words**
- **Documents can have other attributes that should be considered during ranking**

© 2017, Jamie Callan

Other Evidence: The Vector Space

How are query-independent document features handled?

- E.g., Page Rank, spam score, difficulty, …

In the vector-space all vectors have the same dimensions

- But, query-independent features don't occur in the query vector

Solution: Embed the vector space score in a utility function

w_{vsm} × Sim (query, $document_i$) +
$w_{pagerank}$ × PageRank ($document_i$) +
w_{spam} × SpamScore ($document_i$) +
$w_{difficulty}$ × DifficultyScore ($document_i$)

In other words … go outside of the vector space

© 2017, Jamie Callan

Other Evidence: BM25

Can BM25 handle non-text features such as PageRank?

- Model a document as consisting of text (T) + other features (F)

$$\begin{aligned} p(R \mid d) &= p(R \mid d_T, d_F) \\ &\propto \text{BM25}(d_T) + \sum_{d_i \in d_F} \log \frac{p(d_i \mid R)}{p(d_i \mid \bar{R})} \\ &\propto \text{BM25}(d_T) + \sum_i w_i F_i(d_i) \end{aligned}$$

Use whatever features $F_i(d_i)$ and weights w_i you want

- The model allows them, but provides no guidance

(Robertson & Zaragoza, 2007)

© 2017, Jamie Callan

Other Evidence: Indri

A uniform p(d) is common … but, can we do better?

- **Other ways that p(d) might be calculated**
  - Based upon Page Rank
  - Based upon spam score
  - Based upon URL depth
  - …

$$p(d \mid q) \propto p(q \mid d)\, p(d)$$

© 2017, Jamie Callan

Other Evidence: Calculating Priors

Suppose the goal is to set p(d) based on URL depth

- Shallow pages are more likely to be high value pages
- Home pages are usually nearer to the root of the web site

A maximum likelihood estimate for a prior based on url depth

- Acquire a dataset of old queries and clickthrough data

$$p_{priorDepth}\left(depth(url)=n\right)=\frac{\sum_{d\in D}\left(depth\left(d.url\right)=n\right)\&\, clicked(d)}{\sum_{d\in D} depth\left(d.url\right)=n}$$

A similar approach works for PageRank and other evidence

© 2017, Jamie Callan

Other Evidence: Different Approaches to Priors

Query Likelihood and KL Divergence are similar

…until priors are introduced

- **Query likelihood** $p(q \mid d) \propto \log p(d) + \sum_{q_i \in Q} \log p(q_i \mid d)$
  - Expressed in Indri as #and (#prior (url) a b c)
- **KL Divergence** $p(q \mid d) \propto \log p(d) + \frac{1}{|Q|} \sum_{q_i \in Q} \log p(q_i \mid d)$
  - Expressed in Indri as #and (#prior (url) #and (a b c))
- **On long queries, priors have a much larger effect on the KL divergence model than on the query likelihood model**

7

© 2017, Jamie Callan

Other Evidence: Are Document Priors Important?

Document priors are a convenient way of introducing query-independent evidence

- E.g., spam score, PageRank, url depth, …

Run	MAP	P@10
No prior	0.0647	0.1920
Spam	0.0745	0.2720
PageRank	0.0502	0.1820
Url	0.0657	0.2620

Perhaps better theory than in the vector space and Okapi

- But … similar effects can be achieved with those models

(Nguyen and Callan, 2011)

8

© 2017, Jamie Callan

Summary

Know how these are supported by each retrieval model

9 © 2017, Jamie Callan

For Additional Information

- M. Bendersky, D. Fisher, and W. B. Croft. UMass at TREC 2010 Web Track: Term dependence, spam filtering, and quality bias. In *TREC 2010 Conference Proceedings (TREC 2010).* 2011.
- D. A. Metzler, Beyond bags of words: Effectively modeling dependence and features in information retrieval. PhD dissertation, University of Massachusetts. 2007.
- D. Nguyen and J. Callan. Combination of evidence for effective web search. In *TREC 2010 Conference Proceedings (TREC 2010).* 2011.
- S. Robertson and H. Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4). 2009.
- G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley. 1989.
- S. Walker, S.E. Robertson, M. Boughanem, G.J.F. Jones, K. Sparck Jones. "Okapi at TREC-6: Automatic ad hoc, VLC, routing, filtering, and QSDR. TREC-6 Proceedings. 1997.

10 © 2017, Jamie Callan