Language Model Document Priors based on Citation and Co-citation Analysis

H. Zhao, X. Hu and J. Huang
College of Computing & Informatics, Drexel University, USA

Introduction

- Rising interest in combining bibliometrics and information retrieval (IR).
  - White, 2007; Mutschke, 2011; Mayr, 2013
- This paper studies incorporating evidence derived from citation and co-citation analysis into a formal IR model.

Background

- Use citation in information retrieval
  - Textual similarity correlates citation similarity (Salton, 1963)
  - “Boomerang” retrieval strategy (Larsen, 2004)
  - Contextualization approach, (Norozi, 2010)
- Language model document priors
  - Document priors are found helpful in task such as entry page finding (Kraaij, 2002).
  - Previous studies have tried citation count (Meij, 2007), document length (Blanco, 2008), document quality (Zhou, 2005), document type (Sørensen, 2012), linkage citation context (Yin & Huang, 2011) etc.

Methodology

- Language model for IR

$$P(D|Q) \mu P(Q|D)P(D)$$

- P(D) is generally assumed to be uniform, thus ignored.
- We go beyond this uniform assumption by using citation and co-citation analysis results to derive the prior probability of a document being relevant, P(D).
  - Cited count
  - Citation induced paper PageRank score
  - Co-citation cluster

The iSearch Dataset

- Basic facts
  - 18,443 book MAchine-Readable Cataloging (MARC) records (BK), 291,246 articles metadata (PN) and 291,246 PDF full text articles (PF) (Lykke, 2010).
  - 66 topics with relevance judgments.
- Citation data
  - 259,093 PNs and PFs are cited by other PNs and PFs at least once. **We used this subset.**

Document Priors and Their Estimation

- Cited Count Prior

$$P_{citedcount\text{-}mle}(D) = \frac{C_i}{\sum_{k=1}^{N} C_k}$$

- PageRank Prior

$$P_{pagerank\text{-}mle}(D) = \frac{PR_i}{\sum_{k=1}^{N} PR_k}$$

- Both are estimated in three ways:
  - Direct Maximum Likelihood Estimation (MLE)
  - MLE with logarithm smoothed value
  - 10 binned estimation

Document Priors and Their Estimation

- Co-citation Cluster Prior
  - Construct the co-citation graph
    - 259,093 vertices and 33,888,861 edges
  - Use Graclus's Normalized Cut algorithm to partition the graph into 10 clusters

$$P_{cocited}(D) = \frac{\#rel.doc.bin_i}{\#doc.bin_i} / \frac{\#doc.bin_i}{\#total.num.docs}$$

- Five-fold cross-validated estimation over the 57 valid topics.

Experiments

- Baseline
  - Query likelihood model with Jelinek-Mercer smoothing: $\lambda = 0.7$
- LM with priors
  - Indri query: `#(combine #prior( PRIOR ) query terms)`

Results

	MAP	P@10	nDCG	BPREF
baseline-noprior	0.1152	**0.1474**	**0.3134**	**0.3079**
citedcount-mle	0.099	0.1351	0.2825	0.2846
citedcount-log-mle	0.1092	0.1439	0.3046	0.3005
citedcount-bin10	0.1139	0.1452	0.3103	0.2943
pagerank-mle	0.1036	0.1386	0.2972	0.2941
pagerank-log-mle	0.1072	0.1421	0.3031	0.2989
pagerank-bin10	0.1137	0.1434	0.3099	0.2969
cocited-bin10	**0.1155**	0.1397	0.3122	0.3013

Conclusion

- Overall, the result is not promising.
- Across different kinds of priors,
  - logarithm smoothed estimation is better than not;
  - binned estimation is better than MLE estimation.
- Possible reasons for no significant improvement
  - Num. of relevant docs in the relevance judgments is relatively small, making estimation difficult.
  - Performance of document priors could be task/query dependent. Need to do query by query analysis of the results.

- More details are on the paper
  - http://ceur-ws.org/Vol-1143/paper4.pdf

- Thank you!