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

Haozhen Zhao and Xiaohua Hu

College of Computing & Informatics, Drexel University, USA
{haozhen.zhao,xh29}@drexel.edu

Abstract. Citation, an integral component of research papers, implies certain kind of relevance that is not well captured in current Information Retrieval (IR) researches. In this paper, we explore ingesting citation and co-citation analysis results into IR modeling process. We operationalize on going beyond the general uniform document prior assumption in language modeling framework through deriving document priors from papers citation counts, citation induced PageRank and co-citation clusters. We test multiple ways to estimate these priors and conduct extensive experiments on the iSearch test collection. Our results do not suggest significant improvements of using these priors over no prior baseline measured by mainstream retrieval effectiveness metrics. We analyze the possible reasons and suggest further directions in using bibliometric document priors to enhance IR.

1 Introduction

Recent years have seen growing interests in combining bibliometrics and information retrieval (IR), the two major specialties of information science [23]. White proposed a synthesis of the two under Sperber and Wilson's relevance theory, leading to a novel Pennant visualization for accessing literature [24]. Extensive researches have been carried on leveraging the inherent regularity and dynamics of bibliographical entities in scientific information spaces to improve search strategies and retrieval quality [15,13].

We participate in this line of inquiry by studying incorporating evidences derived from citation and co-citation analysis into a formal IR model. Though the importance of citation in assisting researchers to access literature is self-evident, there are not many studies on incorporating them into formal retrieval models. Mainstream IR modeling researches generally center around *term* weighting, smoothing, matching, etc. Still it is possible to ingest bibliometric insights into formal IR models if they are conceptualized as query independence evidences or static features [3]. We adopt here the language modeling framework to investigate whether including citation and co-citation information as document prior probabilities of being relevant to queries improves retrieval effectiveness. (See Section 3.2 for details) We used three kinds of data to estimate document priors: (1) Paper's citation count, (2) Paper's PageRank induced from citation relationships, and (3) Paper co-citation clusters. We compare each approach in

terms of general retrieval effectiveness measurements with extensive experiments on the iSearch test collection[1].

In Section 2, we review related work as the context of our work. Section 3 details on our retrieval model, experiment setup, and the document priors we choose and their estimation methods. Section 4 reports our experiment results and discussion. Section 5 concludes the paper with future directions.

2 Related Work

2.1 Using citation in information retrieval

Garfield initiated the idea of creating citation indexes for scientific articles [6]. Smith reviewed early researches in using citation relations in information retrieval [21]. Salton found out that textual similarity correlated with citation similarity and proposed using terms from bibliographic citation documents to augment original document representation [19]. Larsen studies the "boomerang" effect, which is to use frequently occurring citations in top retrieval result to query against citation indexes for relevant documents [10]. Yin et al. studied linearly combining content score and link score to improve biomedical literature retrieval [25]. For the iSearch test collection, Norozi et al. experimented with a contextualization approach to boost document scores with their random walked neighborhood documents over the in-link and out-link citation network [16]. Document co-citation, as a methodology was proposed by Small, is mostly used in revealing scientific information structure [20]. In this paper, we explore using document co-citation clusters for document prior estimation.

2.2 Language Model Document Priors

Prior information is shown to be useful in certain Web search tasks, e.g. entry page finding [9]. The language model provides an elegant and principled framework to include document priors. Previous studies have used citation counts [14], document length [1], document quality [27], URL type [18,9] and so on as language document priors. For the iSearch collection, there are studies that use the document type as prior [22], as well documents matched with disambiguated query terms[11], in which documents get a higher prior probability of relevance if they match disambiguated query terms. We further this line of study and introduce document co-citation analysis in estimating document priors and compared it with other methods.

3 Methodology

3.1 Dataset

We use the iSearch test collection as our test collection. The iSearch collection was created by the iSearch team. It consists of 18,443 book MAchine-Readable

[1] `http://itlab.dbit.dk/~isearch/`

Cataloging (MARC) records (BK), 291,246 articles metadata (PN) and 291,246 PDF full text articles (PF), plus 3.7 million extracted internal citation entries among PN and PF. 66 topics drawn from physics researchers' real information needs with corresponding relevance judgment data also come with the collection [12]. Previous study has shown that the iSearch collection is appropriate to informetric analysis [7]. Of all the PN and PF documents, 259,093 are cited at least once, which is chosen as the subset for our experiment for reducing citation sparsity consideration. We index them with Indri[2]. Following the best practice in [22], we used the SMART stopword list and Krovetz stemming method. Accordingly, we removed documents not in our index from the relevance judgement files. Then we filter out topics without any relevant documents in the relevance judgement data, resulting 57 valid topics out of the original 66 topics (topic 5, 6, 15, 17, 20, 25, 42, 54, 56 are excluded). We used only the "search_terms" field of the topics as our queries.

3.2 Retrieval Model

We use language model as our IR modeling framework. In particular, we choose the query-likelihood language model [4]. In this model, the relevance of a document D to a query Q is modeled as how likely a user would pose such a query for this document, $P(D|Q)$. Using Bayesian rule, $P(D|Q)$ can be rewritten as:

$$P(D|Q) \propto P(Q|D)P(D), \tag{1}$$

which is easier to be estimated and implemented in IR systems. Much work has been done in finding effective ways to smooth $P(Q|D)$, but generally document prior $P(D)$ is assumed to be uniform thus not affecting the ordering of the retrieval results therefore being ignored [26]. Here we go beyond this uniformity assumption by focusing on the estimation of $P(D)$ with citation and co-citation analysis results. We propose three kinds of priors based on citation counts, citation induced paper PageRank and co-citation clusters.

3.3 Document Priors and Their Estimation

Analyzing paper citation and co-citation network of the iSearch dataset, we propose three kinds of document priors: paper citation count, paper PageRank score induced from citation relationships and co-citation clusters. We tested two kinds of prior estimation methods: maximum likelihood estimation (MLE) and binned estimation. For the MLE approach we also tried a logarithm version. We explain here the three kinds of document priors and how to calculate them.

Paper Citation Count Prior In this case, document prior $P(D)$ is directly estimated based on the proportion of the number of times of a paper being cited (C_i) to the total number of times of all papers being cited:

$$P_{\mathrm{citedcount-mle}}(D) = \frac{C_i}{\sum_{k=1}^{N} C_k}, \tag{2}$$

[2] `http://www.lemurproject.org/indri.php`

and the logarithm version:

$$P_{\text{citedcount}-\log-\text{mle}}(D) = \frac{\log(C_i)}{\sum_{k=1}^{N}\log(C_k)}. \quad (3)$$

Paper PageRank Prior We use the internal citation structure of the iSearch test collection to calculate the PageRank value for all the papers in our index. The PageRank value of a given paper d is:

$$\text{PageRank}(d) = \lambda \sum_{x \in D_{*\to d}} \frac{\text{PageRank}(x)}{|D_{d\to *}|} + \frac{1-\lambda}{N}, \quad (4)$$

where $D_{*\to d}$ and $D_{d\to *}$ denotes papers citing d and cited by d respectively, N is the total number of papers in the collection. $\lambda = 0.85$ is called damping factor [17]. Let PR_i be the PageRank score of paper i, then document PageRank prior using MLE is:

$$P_{\text{pagerank}-\text{mle}}(D) = \frac{\text{PR}_i}{\sum_{k=1}^{N}\text{PR}_k}, \quad (5)$$

and the logarithm version:

$$P_{\text{pagerank}-\log-\text{mle}}(D) = \frac{\log(\text{PR}_i)}{\sum_{k=1}^{N}\log(\text{PR}_k)}. \quad (6)$$

Paper Co-citation Cluster Prior In this case, documents get prior probabilities based on the cluster they belong to. We calculated the document co-citation counts and compiled all the co-citation among the indexed papers, resulting a weighted undirected graph with 259,093 vertices and 33,888,861 edges, with edge weights being the number of times two papers are cited together. We then use the graph clustering software Graclus[3] to cluster the document co-citation network. Graclus provides two clustering algorithms, Normalized Cut (NCT) to minimize the sum of edge weights between clusters and Ratio Association (ASC) to maximize edge density within each clusters [5]. We tried both algorithms and decided to use NCT here because with ASC, most papers are easily clustered into one huge cluster, preventing effective prior estimation.

In the co-citation binned estimation method, the probability a document d from a given bin is given by:

$$P_{\text{cocited}}(D) = \frac{\#\text{ relevant documents of } a \text{ bin}}{\#\text{ documents of } a \text{ bin}} \Big/ \frac{\#\text{ documents of } a \text{ bin}}{\#\text{ total number of documents}}. \quad (7)$$

We used a cross validation method to estimate $P(D)$ in bins. We first order the 57 topic randomly and divide them into 5 folds (11, 11, 11, 12, 12). Then at each round we use 4 folds to estimate the $P(D)$, and use the other 1 fold to test with the prior. We rotate 5 rounds, with each fold being testing set once, then we average results in all the testing folds as the final scores.

[3] `http://www.cs.utexas.edu/users/dml/Software/graclus.html`

We also applied binned estimation methods on Citation Count and PageRank priors. We divide all papers into 10 bins and used the aforementioned five fold cross validation approach to geting the final scores. In total, there are 8 runs reported in Table 1

All estimated $P(D)$ values are converted into logarithm values and applied as Indri prior files and combined with the index using `makeprior` application of Indri. During the retrieval process, they are applied to query terms according to the Indri Query Syntax `#combine(#prior( PRIOR ) query terms)`.

4 Experiment Results and Discussion

With the baseline no prior setup, we extensively tested Jelinek–Mercer (JM) smoothing with $\lambda \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99\}$, Dirichlet prior smoothing with $\mu \in \{100, 500, 800, 1000, 2000, 3000, 4000, 5000, 8000, 10000\}$, and two-stage smoothing with $\{\lambda \times \mu\}$. We find JM smoothing with $\lambda = 0.7$ performs top almost on all the four metrics we chosen. Therefore, we choose it as our retrieval model setting for the reporting baseline and other runs. For each run, we report four mainstream retrieval effectiveness measurements: Mean Average Precision (MAP), Precision at 10 (P@10), nDCG [8] and BPREF[2].

	MAP	P@10	nDCG	BPREF
`baseline-noprior`	0.1152	**0.1474**	**0.3134**	**0.3079**
`citedcount-mle`	0.0990	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

Table 1. Retrieval performance using different document priors and estimation methods compared with baseline using no prior. The best overall score is shown in bold.

Table 1 shows our results in different setups. We can see that the overall effectiveness of applying document priors based on citation counts, PageRank and co-citation clusters is limited. The only marginal improvement over the baseline happens in `cocited-bin10` on MAP. But we can still see difference across priors: overall, logarithm smoothed estimations are better than non-smoothed; binned estimations perform better than MLE estimation.

There are several possible reasons for our results. First, our relevant documents set is relatively small. The total number of relevant documents in our subset of the iSearch test collection qrels is 964, of which there are 863 distinct documents. Though that averages to 17 (964/57) relevant documents for each

topic, more than half of topics (29) has only 7 or fewer documents judged as being relevant. This may contribute to the underperformance in binned estimation of document priors. Second, our current approach is totally independent to content features, only considering the citation dimension. A better approach may be to combine citation features with content features or to use document priors in a query dependent manner. Third, performance of document priors may depend on the type of search tasks or queries. We need to do query by query analysis and comparison of the document priors performance.

5 Conclusion and Future Directions

In this paper, we explored ways of integrating citation and co-citation analysis results into language model modeling framework as document priors. We test three types of document priors with various ways of estimating them. The overall experiment results do not suggest significant improvements over no prior baseline run. In the future, we plan to test document priors with other bibliographic entities such as authors and journals, and to investigate how to effectively combining different kinds of bibliometric-based priors to enhance IR.

Acknowledgements

We appreciate the iSearch team for sharing the iSearch dataset and the helpful comments from the reviewers.

References

1. Roi Blanco and Alvaro Barreiro. Probabilistic document length priors for language models. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *Advances in Information Retrieval*, number 4956 in Lecture Notes in Computer Science, pages 394–405. Springer Berlin Heidelberg, January 2008.
2. Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, page 25–32, New York, NY, USA, 2004. ACM.
3. Nick Craswell, Stephen Robertson, Hugo Zaragoza, and Michael Taylor. Relevance weighting for query independent evidence. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 416–423, New York, NY, USA, 2005. ACM.
4. W. Bruce Croft and John D. Lafferty. *Language Modeling for Information Retrieval*, volume 13 of *The Information Retrieval Series*. 2003.
5. I.S. Dhillon, Yuqiang Guan, and B. Kulis. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):1944–1957, 2007.
6. Eugene Garfield. Citation indexes for science. *Science*, 122:108–111, 1955.

7. Tamara Heck and Philipp Schaer. Performing informetric analysis on information retrieval test collections: Preliminary experiments in the physics domain. In *14th International Society of Scientometrics and Informetrics Conference ISSI*, volume 2, pages 1392–1400, 2013.
8. Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.
9. Wessel Kraaij, Thijs Westerveld, and Djoerd Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 27–34, New York, NY, USA, 2002. ACM.
10. Birger Larsen. *References and citations in automatic indexing and retrieval systems : experiments with the boomerang effect*. PhD thesis, 2004.
11. Christina Lioma, Alok Kothari, and Hinrich Schuetze. Sense discrimination for physics retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 1101–1102, New York, NY, USA, 2011. ACM.
12. Marianne Lykke, Birger Larsen, Haakon Lund, and Peter Ingwersen. Developing a test collection for the evaluation of integrated search. In Cathal Gurrin, Yulan He, Gabriella Kazai, Udo Kruschwitz, Suzanne Little, Thomas Roelleke, Stefan Rüger, and Keith van Rijsbergen, editors, *Advances in Information Retrieval*, number 5993 in Lecture Notes in Computer Science, pages 627–630. Springer Berlin Heidelberg, January 2010.
13. Philipp Mayr and Peter Mutschke. Bibliometric-enhanced retrieval models for big scholarly information systems. In *IEEE International Conference on Big Data (IEEE BigData 2013). Workshop on Scholarly Big Data: Challenges and Ideas*, 2013.
14. Edgar Meij and Maarten de Rijke. Using prior information derived from citations in literature search. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, RIAO '07, pages 665–670, Paris, France, France, 2007.
15. Peter Mutschke, Philipp Mayr, Philipp Schaer, and York Sure. Science models as value-added services for scholarly information systems. *Scientometrics*, 89(1):349–364, 2011.
16. Muhammad Ali Norozi, Arjen P de Vries, and Paavo Arvola. Contextualization from the bibliographic structure. In *Proc. of the ECIR 2012 Workshop on Task-Based and Aggregated Search (TBAS2012)*, page 9, 2012.
17. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-0120, Computer Science Department, Standford University, 1999.
18. Jie Peng, Craig Macdonald, Ben He, and Iadh Ounis. Combination of document priors in web information retrieval. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, RIAO '07, pages 596–611, Paris, France, France, 2007.
19. Gerard Salton. Associative document retrieval techniques using bibliographic information. *J. ACM*, 10(4):440–457, October 1963.
20. H. Small and B. C. Griffith. The structure of scientific literatures i: Identifying and graphing specialties. *Science studies*, pages 17–40, 1974.
21. Linda C. Smith. Citation analysis. *Library Trends*, 30(1):83–106, 1981.
22. Diana Ransgaard Sørensen, Toine Bogers, and Birger Larsen. An exploration of retrieval-enhancing methods for integrated search in a digital library. In *TBAS 2012: ECIR Workshop on Task-based and Aggregated Search*, pages 4–8, 2012.

23. H. D. White and K. W. McCain. Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, 49(4):327–355, 1998.
24. Howard D. White. Combining bibliometrics, information retrieval, and relevance theory, part 1: First examples of a synthesis. *Journal of the American Society for Information Science and Technology*, 58(4):536–559, 2007.
25. Xiaoshi Yin, Jimmy Xiangji Huang, and Zhoujun Li. Mining and modeling linkage information from citation context for improving biomedical literature retrieval. *Information Processing & Management*, 47(1):53–67, January 2011.
26. ChengXiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, April 2004.
27. Yun Zhou and W. Bruce Croft. Document quality models for web ad hoc retrieval. In *CIKM '05 Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 331–332, Bremen, Germany, 2005. ACM.