Using Prior Information Derived from Citations in Literature Search

Edgar Meij & Maarten de Rijke
ISLA, University of Amsterdam
The Netherlands
{emeij, mdr}@science.uva.nl

Abstract

Abstract Researchers spent a large amount of their time searching through an ever increasing number of scientific articles. Although users of scientific search engines prefer the ranking of results according to the number of citations a publication has received, it is never investigated whether this notion of authorativeness could also benefit more traditional and objective measures. Is it also an indicator of relevance, given an information need? In this paper, we examine the relationship between citation features of a scientific article and its prior probability of actually being relevant to some information need. We propose various ways of modeling this relationship and show how this kind of contextual information can be incorporated within a language modeling framework. We experiment with three document priors, which we evaluate on three distinct sets of queries and two document collections from the TREC Genomics track. Empirical results show that two of the proposed priors can significantly improve retrieval effectiveness, measured in terms of mean average precision.

1. Introduction

The body of scientific literature is growing at a staggering rate. Numerous discoveries are disseminated through publications in all fields of modern science. The advent of digital formats and the web have further amplified the availability and accessibility of scientific publications. This has, in turn, enabled scientists worldwide to learn from each other's findings more easily and more rapidly but makes, on the other hand, literature access an increasingly difficult problem. Users in search of specific scientific articles have a plethora of options to choose from. They can go to domain- or publisher-specific search engines, such as Medline or Elsevier's ScienceDirect, or more general search engines for scientific literature such as Citeseer, Citebase, and Google Scholar. The result pages of the systems in the latter category indicate that particular citation features are important to the users (e.g., how many times a particular document has been cited). So, citation features can enhance retrieval *quality* as perceived by end-users, by reflecting a publication's popularity (Redner, 1998). Amento, Terveen et al. have shown that, within a web setting, this is indeed the case (2000). We follow their proposition of separating quality and relevance and pose the question whether applying citation-based features also improves retrieval effectiveness in more objective measures such as mean average precision? Can we use some measure of authorativeness, based on citation features, to improve scientific literature access? To address these questions, we zoom in on publications within the biomedical domain. We explore the relationship between relevance of documents for a given set of queries and the number of citations they receive, and we experiment with various ways of modeling this relationship. We show how knowledge about the bibliographic structure of a document collection can easily be incorporated in a retrieval model based on statistical language models. Instead of assuming uniform prior probabilities of documents being relevant, we introduce a bias towards more often-cited documents. We hypothesize that this bias will actually improve retrieval effectiveness in terms of mean average precision (MAP).

For evaluation purposes, we use the collections and relevance judgments made available by the TREC Genomics track (Hersh, Cohen et al., 2005). Since we use two distinct document collections, we also make observations regarding the influence of the size of the document collections on our findings, with the 2006 collection being significantly smaller. Our main contribution is the novel application of using the authorativeness of a document, as measured by the number of citations it receives, as an indicator of relevance.

2. Related Work

Much of the recent related work has focused on utilizing link-based information in a web retrieval setting. Within a web setting, *importance* of documents (web pages) can be captured using hyperlink-based information (Kleinberg, 1999; Page, Brin et al., 1998). Both HITS and PageRank are based on the assumption that a document which is referenced many times by other documents is more important (or *authorative*). When the referring document is authorative itself, the authorativeness of the referred documents increases. Despite the fact that these algorithms generally improve the perceived *quality* of the results of an IR system (Amento, Terveen et al., 2000; Kleinberg, 1999), the actual improvement in *relevance* scores for adhoc search is not proven (Hawking and Craswell, 2001; Hawking, Voorhees et al., 1999).

Citation indexing was introduced in the 1950s as a means to keep track of references that authors put in their bibliographies (Garfield, 1955). It provided a way to analyze the literature and gather data on the ``impact'' of authors, organizations, countries, and journals, as well as assessing particular areas of research activity and publication. There has been much debate on the relevance of the published *Impact Factor*, with a general conclusion of it being a ``bibliometric tool with limited explanatory power'' (Brody, 1995; Dong, Loh et al., 2005; Gowrishankar and Divakar, 1999; Opthof, 1997). To the best of our knowledge, this is the first time anyone has attempted to verify the assumption that authorativeness, as measured through citation features, is indeed a contributing factor not only to quality, but to relevance in particular — within the setting of a scientific document search task.

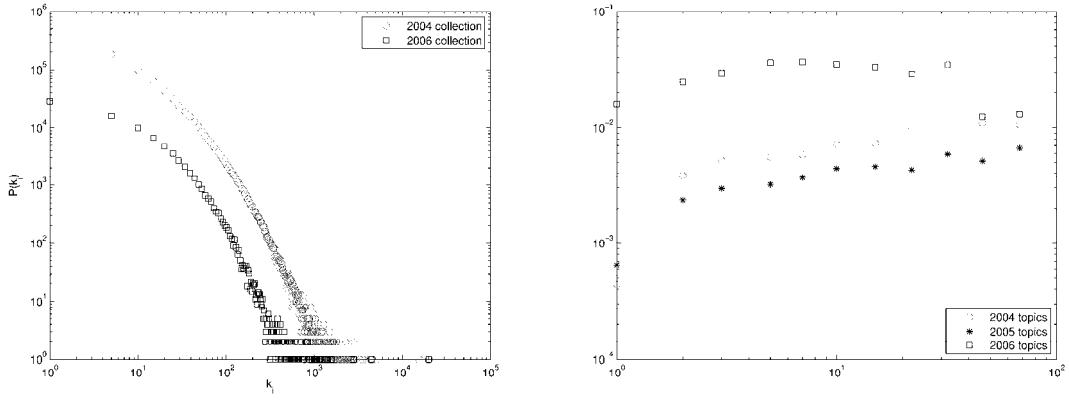
3. Background

3.1 Scale-Free Random Graphs

Barabási and his colleagues were the first to map the connectedness of the web in 1999 (Barabási and Albert, 1999; Barabási, Albert et al., 1999). They noticed that it was a compact network and that the distribution of the numbers of connections of its vertices had an unusual *fat-tailed* form. They proposed two possible causes for the emergence of this power law in the frequency of connectivity: incremental growth and preferential attachment. *Incremental growth* refers to networks that expand continuously by the addition of new nodes, and thus the gradual increase in the size of the network. *Preferential attachment* refers to the tendency of a new node to connect to existing nodes that are already highly connected. These two properties are the building blocks of the proposed Barabási-Albert (BA) model. In this model, a vertex is introduced in the network at each timestep, with *m* edges. The probability that a new node *k* is connected to node *i* is $\Pr(k_i) = k_i \Big/ \sum_j k_j$ (1). Thus, the new vertex is more probable to connect to already highly connected or *popular* vertices (which is why this model is also referred to as the “rich get richer” model (Barabási and Albert, 1999). Typically, the majority of vertices have a small number of edges and a few highly connected ones are authorative and function as hubs. When $k \to \infty$, the probability that node *i* interacts with *k* other nodes decays as $\Pr(k) \sim k^{-\gamma}$ and thus has no natural scale, making the local connectivity distribution *scale free*. When we assume the discrete variable k_i to be continuous, it can be shown that the rate $\delta k_i / \delta t$ at which k_i changes over time equals $k_i / 2t$. Taking the boundary condition that node *i* was added at time t_i, with k_o outlinks into account, the solution to this equation is of the form:

$$k_i = k_o \sqrt{\frac{t_i}{t}}, \qquad (2)$$

which yields the expected indegree at time *t* (Dorogovtsev and Mendes, 2002; Dorogovtsev and Mendes, 2003). Later, we will compare this *expected* number of received citations with the actual numbers and use the normalized difference as a document prior.



a b

Figure 1: Distribution of received citations of all documents in the collections (a) and the number of received citations versus prior probability of relevance (b), both on a log-log plot.

3.2 Language Modeling

In our modeling and experimentation we adopt a generative language modeling setting. We estimate a multinomial unigram language model for each document in the collection and for any given query we rank the documents with respect to the likelihood of a document generating the query. To account for data sparseness, we interpolate the this likelihood using Jelinek-Mercer smoothing (Hiemstra, 2001; Zaragoza, Hiemstra et al., 2003; Zhai and Lafferty, 2001). We assume the query terms to be independent, and use a linear interpolation of a document model and a collection model to estimate the probability of each query term t in query q:

$$P(q \mid d) \propto P(d) \cdot \prod_{t \in q} (1-\lambda) P(t \mid D) + \lambda P(t \mid d), \tag{3}$$

where d is a document and D is the collection. We need to estimate three probabilities: the probability of observing a term in a document, $P(t|d)$; the probability of observing the term in the collection, $P(t|D)$; and the prior probability of the document, $P(d)$. The first two are estimated using regular maximum likelihood estimates (Croft and Lafferty, 2003), while the document priors provide us with an elegant mechanism for incorporating external sources of evidence. Usually, all documents are considered to be equally relevant and a uniform document prior is chosen: $P_{uniform}(d) = 1/|D|$, where $|D|$ denotes the total number of documents in the collection.

4. Experimental Design

To assess the effect of introducing document priors based on bibliometric information, we perform several experiments, for which we use the retrieval model introduced in the previous section. In this section we detail our test collections and describe the sources we accessed to gather the required bibliometric data.

Collections PubMed is a service of the U.S. National Library of Medicine (NLM) that indexes Medline, the online database of the NLM. The TREC Genomics track is concerned with information needs and documents of a biomedical nature — it uses a subset of Medline as its document collection(s) (Hersh, Cohen et al., 2005). It is customary within this track to perform extensive domain-specific morphological preprocessing of the documents; see, e.g., (Huang, Ming et al., 2005). As we want our findings and conclusions to be as generic as possible, we have not performed any of these forms of specialized preprocessing for the experiments we present here. We use the topics and collections from the TREC Genomics track (2004 - 2006); the 2004

TREC Genomics document collection was used for both the 2004 and 2005 topic sets and contains 4,591,098 abstracts of biomedical articles, with the 2004 and the 2005 topic sets each containing 50 queries. For the 2006 Genomics track a new collection was introduced, consisting of 162,259 full-text biomedical articles and 28 topics. Additionally, in 2006 passage retrieval was introduced as the new task. Participating systems were not only judged on the document level, but also on passage and aspect levels. Again, for our findings to be as generic as possible, we only consider relevance judgments at the document level.

Bibliometric Data PubMed Central[1] is the NLM's free digital archive of biomedical and life sciences literature. One of the services it provides is an interface to a citation index of the records in MedLine and we use this bibliometric data as a basis for our experiments. We retrieved a total of 1,048,423 citing publications for the 2004 collection, whereas we found 84,800 for the 2006 collection. When an article does not have any bibliometric information, it means it either has not received any citations, or that PubMed Central has not indexed it. Given the rate at which new articles are published and included in MedLine (Yoo, 2006), the former situation is not unthinkable and we believe that the amount of citations we have acquired is representative for the task at hand. The number of received citations per document does not follow a normal distribution. Indeed, Figure 1a displays the probability distribution from Equation 1 on a log-log plot for both collections and the fat-tailed curves typical of a power-law distribution are clearly visible.

5. Estimating Bibliometrical Priors

To investigate whether bibliometric information can be a useful indicator of the prior probability of a document's relevancy, we plot the distribution of number of received citations versus relevancy (Figure 1b). From this figure it is clear that a relationship between the number of received citations and the prior probability of a document being relevant exists. We now introduce three document priors into Equation 3, which are all based on citation features.

Scale Free – Maximum Likelihood Estimation This document prior is based on Equation 1 and captures the relative authoritativeness of a given document. It estimates the distribution of received citations and normalizes the raw counts: $P_{mle}(d) = k_i(d) \Big/ \sum_{d' \in D} k_i(d')$, where $k_i(d)$ denotes the number of citations document d receives.

Scale Free There is obviously a temporal aspect to the number of citations a document receives. Our intuition is that this number is not only dependent on the quality, but also on the age of a particular document, in line with the ideas underlying the BA model introduced in Section 3.1. Recent publications are less likely to have a larger number of received citations, simply because it takes time for a paper to get noticed and thus cited. Following this idea put forward by (Hauff and Azzopardi, 2005), we calculate the expected number of received citations $k_i^{\exp}(d)$ for every document, using Equation 2. Since we also have the actual number of received citations $k_i^{\text{act}}(d)$ available, we compare the two and calculate the difference $k_i^{\text{act}}(d)$ - $k_i^{\exp}(d)$. We normalize the differences and use the resulting values as prior P_{sfn}.

Bins We use an equal width interval discretization with 6 bins for this prior (this number is based on the heuristic reported by (Dougherty, Kohavi et al., 1995)). Every bin represents a range of received citations, with a value for the prior P_{bins}. We use leave-one-out cross-validation to estimate the value of the prior: for every topic set, we use the relevance judgments of one topic to estimate the optimal prior value for every bin and exclude that particular topic from the evaluation. We exclude every topic exactly once and take the mean of the resulting scores as the final score.

[1] PubMed Central, http://www.pubmedcentral.gov

	2004 topic set (λ = 0.05)		2005 topic set (λ = 0.2)		2006 topic set (λ = 0.15)	
	MAP	Change	MAP	Change	MAP	Change
Baseline	0.2374		0.2115		0.2728	
P_{mle}	**0.2549**	+7.37 %***	**0.2362**	+11.68 %***	0.2726	-0.07 %
P_{bins}	0.2379	+0.21 %	0.2121	+0.28 %*	**0.2860**	+4.84 %***
P_{sfn}	0.2468	+3.96 %	0.2106	-0.43 %	0.2731	+0.11 %

Table 1: Experimental results of different methods and topic sets (best scores in boldface).

6. Results and Discussion

In this section we elaborate on the results presented in Table 1. These results were obtained by estimating the various priors on the collection associated with the reported topic set. As a baseline we use the query-likelihood model from Equation 3 with a uniform document prior. We fix λ for the various prior experiments, based on the best performing baseline run per topic set. We use a Wilcoxon signed-rank test and look for improvements at significance levels 0.95(*), 0.99(**), and 0.999(***). Note that the improvements presented here cannot be attributed to a low baseline — our baseline scores are well above the median of all participants' scores at the TREC Genomics tracks 2004 - 2006. The scale-free network prior P_{sfn} does not lead to any significant improvements. It only improves results on the 2004 topics slightly, but not significantly. For the 2006 topic set, this prior roughly helps as much as it hurts retrieval effectiveness. It seems that the modeling assumptions as proposed by (Barabási and Albert, 1999) do not hold true for these collections, despite the evidence in favor (Redner, 1998). When we turn to the other bibliometric priors, we note the improvement in terms of MAP of up to 11% on the 2005 topic set, using the maximum likelihood estimation document prior. This prior yields similar results on the 2004 topic set, but does not perform well on the 2006 topics. The difference in retrieval effectiveness could be caused by the difference in size between the collections, but a more plausible explanation can be found in Figure 1b, which shows a less clear relationship between prior probability of relevance and the number of citations.

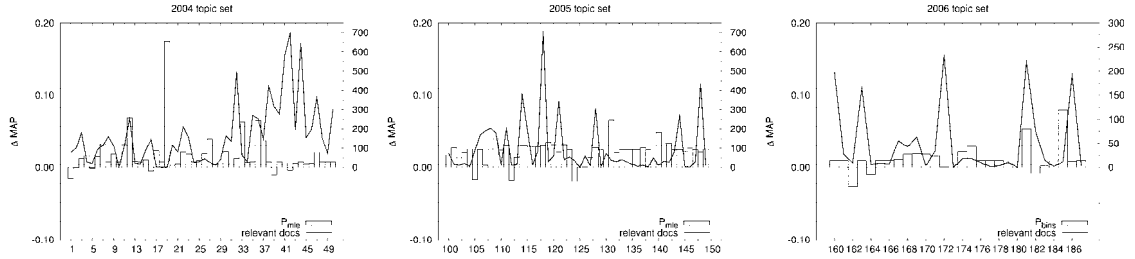


Figure 2: Per-topic breakdown of improvement over baseline of the best performing prior per topic set (bars), with the amount of relevant documents per topic included (lines).

The fact that the 2006 collection is much smaller also provides an explanation as to why the binned document prior performs better here. The proposed cross-validation method is a way to estimate an underlying probability distribution, which is more robust to smaller datasets. This does not explain the lower scores for this prior on the 2004 collection however. To gain further insight into this phenomenon we also investigated the per-topic scores per prior. The resulting plots can be found in Figure 2. The individual topic scores for the P_{bins} prior actually represent the scores of all the other topics, when trained on that particular topic. From this figure it is clear that the P_{mle} prior improves nearly all 2004 and 2005 topics, but has little effect on the 2006 topics. Similar behavior can be observed in the opposite direction for the P_{bins} prior. We believe that the shift in task from the 2005 to the 2006 TREC Genomics track may have some influence. The estimation of the value for these priors may also be influenced by the distribution of the number of relevant documents per topic. When we superimpose the number of relevant documents on the earlier presented per-topic difference in retrieval effectiveness (also in Figure 2), we see a slightly

different distribution between the 2004 and 2006 collections. Contrary to the lessons learned from the TREC Web track, which suggest that inlink-type priors work best for named or home page finding tasks — with very few relevant documents — this indicates that this proved not to be the case with the current task, topics, and collections.

7. Conclusion

The vastly growing number of scientific publications calls for additional ways to rank documents when faced with a particular information need. We have shown how this kind of bibliometric information regarding the authorativeness of a scientific article can easily be integrated into a language modeling framework and have presented various ways of modeling the relationship between relevance and citedness. We have shown that the prior probability of a scientific article being relevant can, to some degree, be captured by using citation features. This relationship is especially visible on a larger document collection. Two of the specific document priors we propose are based on the BA model as introduced in Section 3.1. One of these, P_{mle}, yields consistent and statistically significant improvements on the 2004 collection, but slightly degrades when applied to the much smaller 2006 collection. The third prior, P_{bins}, fails to make significant improvements on the larger collection, but does improve results on the smaller 2006 collection in part because of the way the relevant documents are distributed among the topics.

References

Amento, B., L. Terveen, et al. (2000). Does ``authority" mean quality? Predicting expert quality ratings of Web documents. *SIGIR '00*.

Barabási, A.-L. and R. Albert (1999). Emergence of Scaling in Random Networks. *Science* 286(5439).

Barabási, A.-L. a., R. e. Albert, et al. (1999). Diameter of the world-wide web. *Nature* 401: 130--131.

Brody, S. (1995). Impact factor as the best operational measure of medical journals. *Lancet* 346.

Croft, W. B. and J. Lafferty (2003). *Language Modeling for Information Retrieval*.

Dong, P., M. Loh, et al. (2005). The ``impact factor" revisited. *Biomedical Digital Libraries* 2(7).

Dorogovtsev, S. N. and J. F. F. Mendes (2002). Accelerated growth of networks. *CoRR*.

Dorogovtsev, S. N. and J. F. F. Mendes (2003). *Evolution of Networks: From Biological Nets to the Internet and WWW*, Oxford University Press.

Dougherty, J., R. Kohavi, et al. (1995). Supervised and Unsupervised Discretization of Continuous Features. *International Conference on Machine Learning*: 194-202.

Garfield, E. (1955). Citation Indexes for Science. *Science* 122.

Gowrishankar, J. and P. Divakar (1999). Sprucing up one's impact factor. *Nature* 401(6751).

Hauff, C. and L. Azzopardi (2005). Age Dependent Document Priors in Link Structure Analysis. *ECIR'05*.

Hawking, D. and N. Craswell (2001). Overview of the TREC-2001 Web Track. *TREC-9*.

Hawking, D., E. Voorhees, et al. (1999). Overview of the TREC-8 Web Track. *TREC-8*.

Hersh, W., A. Cohen, et al. (2005). TREC 2005 Genomics Track Overview. *TREC-14*.

Hiemstra, D. (2001). Using Language Models for Information Retrieval, University of Twente.

Huang, X., Z. Ming, et al. (2005). York University at TREC 2005: Genomics Track. *TREC-14*.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM* 46(5): 604--632.

Opthof, T. (1997). Sense and nonsense about the impact factor. *Cardiovasc Res* 33(1): 1--7.

Page, L., S. Brin, et al. (1998). The PageRank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project.

Redner, S. (1998). How Popular is Your Paper? An Empirical Study of the Citation Distribution. *The European Physical Journal B* 4: 131.

Yoo, I. (2006). Semantic text mining and its application in biomedical domain.

Zaragoza, H., D. Hiemstra, et al. (2003). Bayesian extension to the language model for ad hoc information retrieval. *SIGIR '03*.

Zhai, C. and J. Lafferty (2001). A study of smoothing methods for language models applied to Ad Hoc information retrieval. *SIGIR '01*.