LAMP-TR-146 January 2008
HCIL-2008-01

PageRank without Hyperlinks: Reranking with Related Document Networks

Jimmy Lin

The iSchool
University of Maryland
College Park, Maryland, USA
E-mail: jimmylin@umd.edu

Abstract

Graph analysis algorithms such as PageRank and HITS have been successful in Web environments because they are able to extract important inter-document relationships from manually-created hyperlinks. We consider the application of these algorithms to related document networks comprised of automatically-generated content-similarity links. Specifically, this work tackles the problem of document retrieval in the biomedical domain, in the context of the PubMed search engine. A series of reranking experiments demonstrate that incorporating evidence extracted from link structure yields significant improvements in terms of standard ranked retrieval metrics. These results extend the applicability of link analysis algorithms to different environments.

Publication Date: January 14, 2008
Keywords: document retrieval, biomedical domain, cluster hypothesis, network analysis, TREC genomics track

Please cite as: Jimmy Lin. PageRank without Hyperlinks: Reranking with Related Document Networks. Technical Report LAMP-TR-146/HCIL-2008-01, University of Maryland, College Park, January 2008.

1 Introduction

One of the most important innovations in information retrieval over the past decade has been the development of algorithms that exploit inter-document relationships. In most cases, documents do not exist in isolation—their environments provide an important source of evidence for ranking results with respect to a user's query. This insight is captured in algorithms such as Google's PageRank (Page et al., 1999) and HITS (Kleinberg, 1999), also known as "hubs and authorities". Both have been successful in Web environments, where hyperlinks provide the inter-document relationships. The two algorithms operationalize in different ways the basic idea that a hyperlink represents an endorsement of the target page by the source author.

This paper considers the application of these algorithms to a different type of graph structure. In the absence of manually-created hyperlinks, we argue that related document networks, or networks defined by content similarity, can be treated in the same manner as hyperlink graphs. Experiments show that incorporating evidence extracted from such networks yields statistically significant improvements in document retrieval performance, as measured by standard ranked-retrieval metrics.

Our work focuses on search in the biomedical domain, in the context of the PubMed search engine. PubMed is a large, publicly-accessible Web-based search engine that provides access to MEDLINE, the authoritative repository of abstracts from the medical and biomedical primary literature, maintained by the U.S. National Library of Medicine (NLM). MEDLINE currently contains over 17 million abstracts, covering a wide range of disciplines within the health sciences (broadly interpreted), from biochemistry to public policy.

Whenever the user examines an abstract in PubMed, the right panel of the browser is automatically populated with titles of articles that may also be of interest, as determined by a probabilistic content similarity algorithm (Lin and Wilbur, 2007); see Figure 1 for an example. In other words, each abstract view automatically triggers a related article search: the top five results are integrated into a "related links" panel in the display.[1]

Related article search provides an effective browsing tool for PubMed users, allowing them to navigate the document collection without explicitly issuing queries. Any given MEDLINE abstract is connected to a number of related articles, which are in turn connected to even more related articles, and so on. Thus, any single abstract represents a node in a vast related document network defined by content similarity links. We explore the hypothesis that these networks could be exploited for document retrieval, in the same manner as hyperlink graphs in the Web environment.

2 Why Related Document Networks?

Despite superficial similarities, the analogy between hyperlink graphs of the Web and related document networks in MEDLINE is far from perfect. Hyperlinks are created by humans and represent intentionality. That is, an author links to another Web page because he or she "likes it". Thus, inbound links can be interpreted as votes of confidence with respect to the quality, authority, etc. of the Web page. Algorithms such as PageRank and HITS take advantage of this basic idea. Related document networks, on the other hand, are artificial. Since they are automatically computed by a content similarity algorithm, the networks reflect inherent characteristics of the document collection, i.e., term distributions. Furthermore, the nature of content similarity algorithms means that every document is related to every other one to some degree; thus, we face a thresholding problem when deciding how expansive a related document network might be. Given these important differences, why might one believe that there is worthwhile information contained in the link structure? In this section, we present independent motivation for our hypothesis.

[1] Although MEDLINE records contain only abstract text, it is not inaccurate to speak of searching for articles since PubMed provides access to the full text when available; we use "document" and "article" interchangeably in this paper.

Figure 1: Screenshot of PubMed showing a MEDLINE abstract. The "Related Links" panel on the right is populated with titles of articles that may be of interest.

One source of support comes from the cluster hypothesis in information retrieval (van Rijsbergen, 1979), which is the simple observation dating back several decades that closely associated documents tend to be relevant to the same requests. Another interpretation is that relevant documents tend to occur in clusters. Many researchers have explored and confirmed this hypothesis as a basic property of document collections to varying degrees (Voorhees, 1985). Therefore, the underlying topology of related document networks might provide clues as to where relevant documents might lie in the collection space.

Similar support comes from cognitive psychology. The theory of information foraging (Pirolli and Card, 1999) hypothesizes that, when feasible, natural information systems evolve toward states that maximize gains of valuable information per unit cost. Furthermore, the theory claims that information seekers behave in a manner that is not unlike our hunter-gatherer ancestors foraging in physical space. One basic assumption in information foraging theory is the notion of information patches—the tendency for relevant information to cluster together. An information seeker's activities are divided between those that involve exploiting the current patch and those that involve searching for the next patch—thus, the user is constantly faced with the decision to pursue one or the other activity. These claims can be understood as a different formulation of the cluster hypothesis: relevant documents co-occur in similarity space, and thus the structure of this space is an important consideration in retrieval. Whereas the cluster hypothesis adopts a system-centered view, information foraging theory focuses on human search behavior. Nevertheless, both converge on the same idea.

Additionally, empirical support comes from usage patterns of related article search. A recent analysis of PubMed query logs indicate that searchers click on suggested article titles with significant frequency (Lin et al., 2007). Data gathered during a one week period in June 2007 indicate that approximately 5% of page views in non-trivial user sessions (discarding, for example, sessions that consist of one page view) are generated from users clicking on related article links. Approximately one fifth of all non-trivial user sessions involve at least one click on a related article link. Furthermore, there is evidence of sustained browsing using this feature: the most frequent action following a click on

a related article is another click on a related article (about 40% of the time). Thus, related document networks appear to be an integral part of PubMed searchers' activities—suggesting that characteristics of these networks might provide an important source of evidence for document ranking.

3 Exploiting Link Structure

This work examines two well-known algorithms that exploit link structure to score the importance of nodes in a hyperlink graph such as the Web: PageRank (Page et al., 1999) and HITS (Kleinberg, 1999). We overview both algorithms, but refer the reader to the original articles for details.

PageRank conceptually models a random Web surfer. Given a tireless, idealized user who randomly clicks on hyperlinks (i.e., participates in a random walk), the measure quantifies the fraction of time that is expected to be spent on any given page. Thus, pages with many in-links or moderate numbers of highly-ranked in-links will have high PageRank scores—this is consistent with our intuition of an "important" Web page. The distribution of PageRank scores can be interpreted as the principal eigenvector of the normalized link matrix. As an additional refinement, PageRank incorporates a damping factor, which models the probability that the surfer will randomly jump to another page (thus avoiding link cycles). Typically, PageRank is computed iteratively, and has been empirically shown to converge in surprisingly few iterations, even for extremely large networks.

The HITS algorithm views the hyperlink graph of the Web as containing a set of "authoritative pages" joined together by a set of "hub pages". The task, therefore, is to discover which nodes are hubs and which are authorities from the link structure (i.e., assign a hub and authority score to every node). Operationally, hubs and authorities are recursively defined in terms of each other: a good hub is a page that points to many good authorities, and a good authority is a page that is pointed by many good hubs. This gives rise to an iterative technique for computing hub and authority scores, although Kleinberg provides a theoretical foundation for his formulation in terms of eigenvectors of matrices associated with the hyperlink graph.

4 Experimental Setup

This section describes the setup of our experiments, starting first with an overview of the test collection. We then detail the procedure used in our reranking experiments.

4.1 Test Collection

Evaluations were conducted with data from the TREC 2005 genomics track (Hersh et al., 2005). One salient feature of this TREC evaluation was its use of generic topic templates (GTTs), which consist of semantic types, such as genes and diseases, embedded in prototypical information needs, as determined from interviews with biologists and other researchers. In total, five templates were developed, each with ten fully-instantiated "topics" (information needs, in TREC parlance); examples are shown in Table 1. In some cases, the actual topics deviate slightly from the template structure (in order to accommodate real requests).

The TREC 2005 genomics track employed a ten-year subset of MEDLINE (1994–2003) containing 4.6 million citations, or approximately a third of the entire database at the time it was collected in 2004 (commonly known as the MEDLINE04 collection). A total of 32 groups submitted 58 runs the evaluation. System output was evaluated using the standard pooling methodology for *ad hoc* retrieval, with relevance judgments supplied by an undergraduate student and Ph.D. researcher in biology. No relevant documents were found for one topic, which was dropped in our experiments.

#1	**Information describing standard [methods or protocols] for doing some sort of experiment or procedure.** *methods or protocols:* chromatin IP (Immuno Precipitations) to isolate proteins that are bound to DNA in order to precipitate the proteins out of the DNA
#2	**Information describing the role(s) of a [gene] involved in a [disease].** *gene:* Transforming growth factor-beta1 (TGF-beta1) *disease:* Cerebral Amyloid Angiopathy (CAA)
#3	**Information describing the role of a [gene] in a specific [biological process].** *gene:* COP2 *biological process:* transport of CFTR out of the endoplasmic reticulum
#4	**Information describing interactions between two or more [genes] in the [function of an organ] or in a [disease].** *genes:* HNF4 and COUP-TF I *function of an organ:* function of the liver
#5	**Information describing one or more [mutations] of a given [gene] and its [biological impact or role].** *gene with mutation:* NM23 *biological impact:* tracheal development

Table 1: Templates and sample instantiations from the TREC 2005 genomics track.

4.2 Document Reranking

We evaluated the impact of features extracted from related document networks in a reranking experiment. The starting point was a ranked list containing 40 hits, retrieved with the Java-based Terrier information retrieval platform[2] using the In_expB2 retrieval model (the c parameter was arbitrarily set to 1). Terrier's retrieval algorithm is based on the divergence from randomness framework, discussed in (Amati and van Rijsbergen, 2002). Template instantiations from the genomics track topics were submitted as the queries, with no special processing.

From this ranked list, we constructed several related document networks by varying the number of related document expansions for each hit in the Terrier result set. That is, for each document, we added links to its 5, 10, 15, and 20 most-similar "neighbors". Naturally, adding more related documents results in greater network density, which as we show has a significant impact on results. PubMed's eutil API allowed us to programmatically retrieve the related documents, which we post-processed to eliminate documents not in our corpus. To avoid combinatorial explosion, we did not perform second order expansions (i.e., related documents of related documents), although that is a possibility. The PageRank and HITS algorithms were then applied to these networks, using the implementation in JUNG (Java Universal Network/Graph Framework),[3] an extensible open source toolkit for network analysis. For PageRank, we set the damping parameter to 0.15, a value frequently suggested in the literature.

Scores extracted from the network were combined with Terrier retrieval scores using weighted linear interpolation, controlled by the parameter λ, i.e., weight of λ to Terrier scores, weight of $(1 - \lambda)$ to network scores. We ran three separate sets of experiments, using PageRank scores, HITS authority scores, and HITS hub scores. Note that our simple evidence combination approach is similar in spirit to discriminative ranking algorithms in IR (Nallapati, 2004) and methods employed by commercial

[2] http://ir.dcs.gla.ac.uk/terrier/
[3] http://jung.sourceforge.net/

search engines (although far simpler since our experiments involve only two features). The output of this scoring process is a new ranking of the documents from the original Terrier ranked list.

We evaluated reranked output in terms of three standard ranked-retrieval metrics: precision at 20 documents (P20), relative mean average precision at 20 document (MAP20), and also at 40 documents (MAP40). The cutoffs were selected since the current PubMed interface displays 20 hits per page. Early precision is important in a Web search context, since users in general examine relatively few results. These metrics capture the quality of the first two result pages (since we are reranking 40 documents, P40 is not informative). Finally, note that we measure relative MAP—that is, with respect only to relevant documents contained in the original Terrier ranked list. This modification was made since we were only working with the top 40 retrieved documents; for topics with more than 40 known relevant documents, a perfect score was impossible (thus making the possible score range for each topic different). This was a concern since our test collection contained an average of 95 relevant documents per topic. Computing relative MAP makes averages across topics more meaningful.

5 Results

Results of our reranking experiments combining Terrier and PageRank scores are shown in Figure 2: MAP20 on top, MAP40 in the middle, and P20 on the bottom. Expansion with a different number of related documents is shown as separate lines. The x-axis represents a sweep across the λ parameter space in tenth increments. Thus, the right edge of each line ($\lambda = 1.0$) represents baseline Terrier results (with no contribution from PageRank scores). For all experiments in this paper, we applied the Wilcoxon signed rank test to assess the statistical significance of performance differences. Points that represent significant improvements over the Terrier baseline ($p < 0.05$) are denoted by solid markers.

The graphs confirm that incorporating PageRank scores using our simple combination approach improves ranked retrieval performance, reaching optimal performance between 0.6–0.8 in terms of λ values. Lower values of λ, representing heavier emphasis on PageRank scores, consistently results in below-baseline performance. For many settings of the number of related document expansions, peak performance is significantly better than the baseline. In general, we note that more related article expansions improve performance. It appears that denser networks provide richer input to the PageRank algorithm, and yields a better estimation of a document's importance in the network.

Corresponding graphs for interpolating Terrier scores with HITS authority scores are shown in Figure 3, and for HITS hub scores, in Figure 4. To facilitate comparison between the different methods, we use the same vertical scale for each metric. As in Figure 2, we also apply the convention of denoting statistically significant improvements ($p < 0.05$) with solid markers. The only cases observed were with HITS authority scores in P20; in all other cases, gains were not statistically significant. We note the same general trends with both sets of HITS scores, although they appear to be less valuable than PageRank for document ranking.

To provide context, it is worthwhile to compare our results to previous runs submitted to the TREC 2005 genomics track. As Hersh et al. (2005) report, the best mean average precision for an automatic run (containing 1000 hits per topic) was 0.289; the best precision at 10 was 0.474 (these were two different runs). The mean for all 58 submitted runs was 0.197 in terms of MAP and 0.358 in terms of P10. As a separate experiment, we generated a comparable baseline run (Terrier using the In_expB2 retrieval model): it achieved 0.255 MAP and 0.428 P10. Since we are using Terrier "out of the box" with minimal modification, we naturally did not expect superlative performance—the best-performing runs all involved techniques to address domain-specific terminology, e.g., through query expansion (Huang et al., 2005). Nevertheless, these results confirm that we are starting with a competitive baseline, suggesting that improvements contributed by link analysis are indeed meaningful.

Although a sweep across the λ parameter space allows us to understand relative importance of link analysis and Terrier retrieval scores, it doesn't tell us if optimal values are realistically achievable.

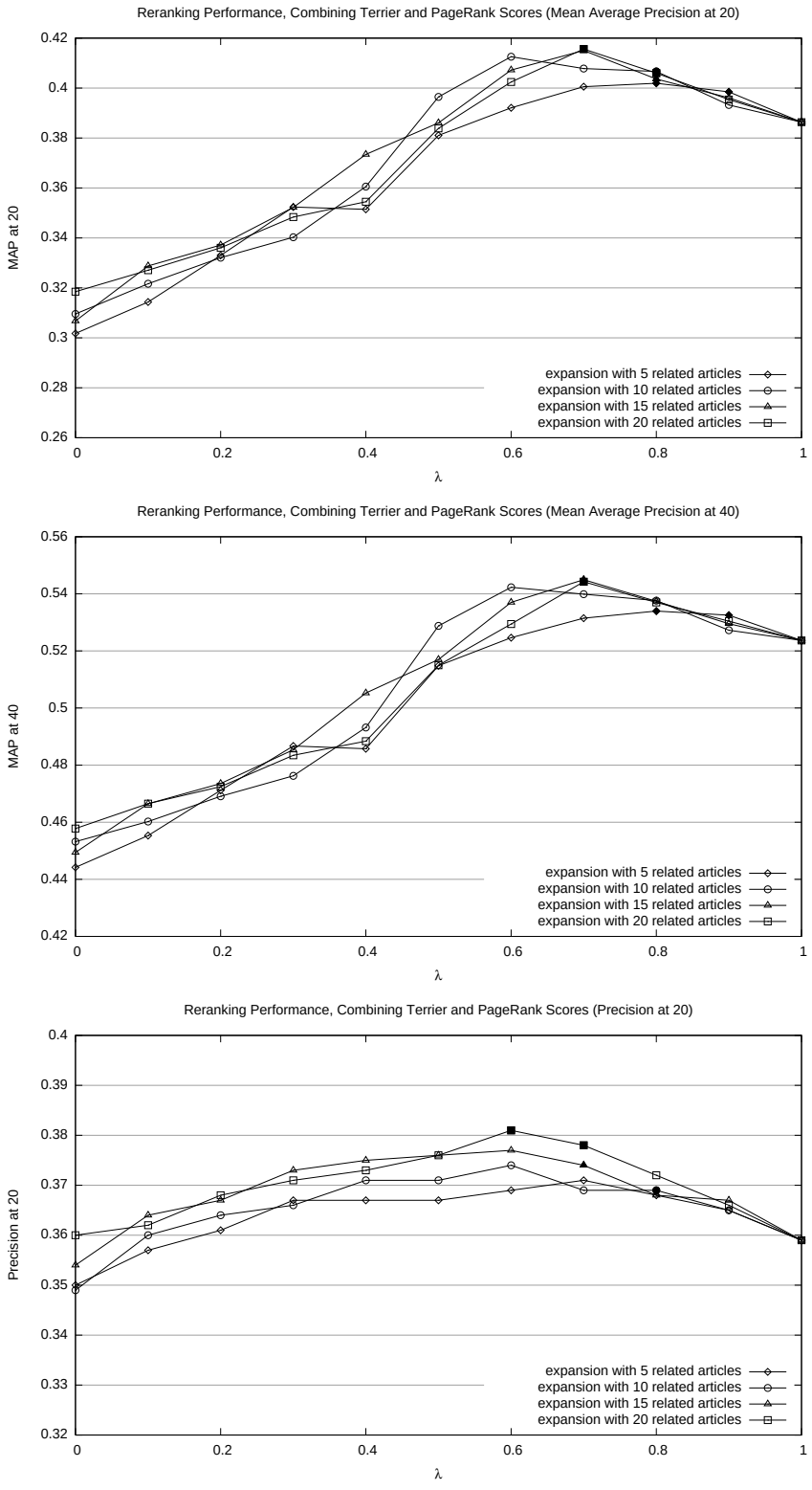

Figure 2: Reranking performance based on interpolating Terrier retrieval scores with PageRank scores: MAP20 (top), MAP40 (middle), and P20 (bottom). Solid markers represent significant improvements.

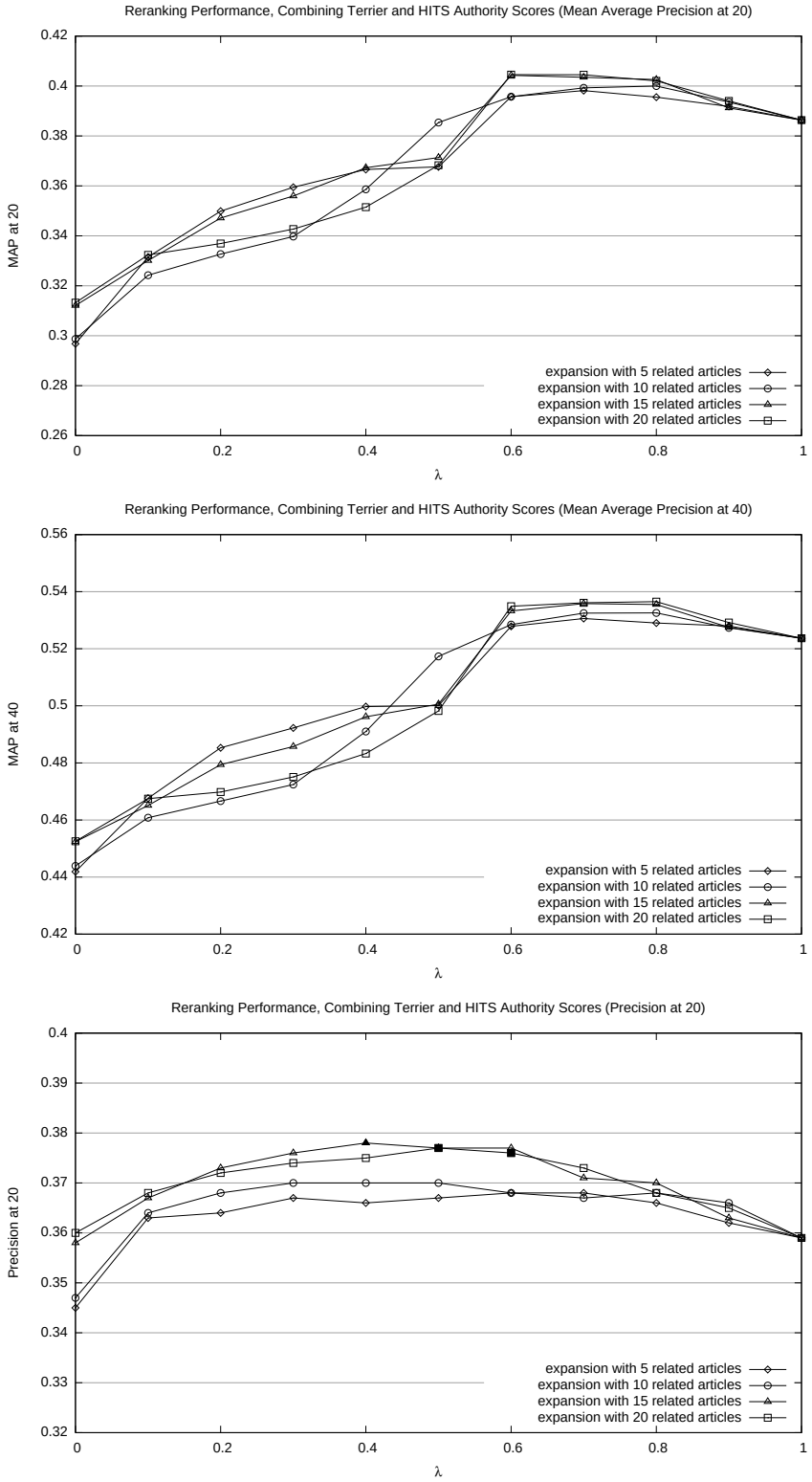

Figure 3: Reranking performance based on interpolating Terrier scores with HITS authority scores.

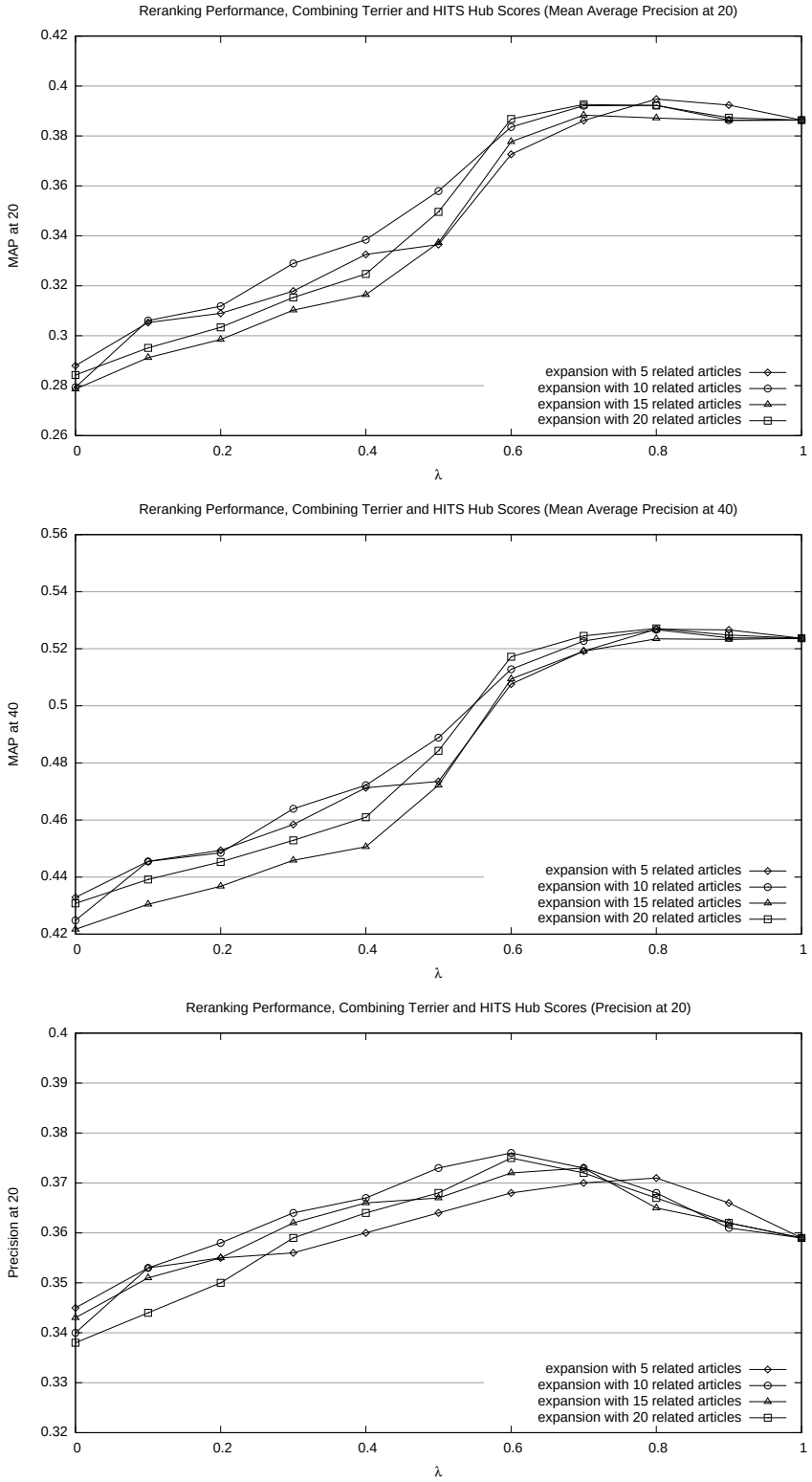

Figure 4: Reranking performance based on interpolating Terrier scores with HITS hub scores.

Tuning on MAP20

	MAP20
Baseline	0.386
Learned ($\lambda = 0.7$)	0.416 (+7.8%)*

Tuning on MAP40

	MAP40
Baseline	0.524
Learned ($\lambda = 0.7$)	0.544 (+3.8%)*

Tuning on P20

	P20
Baseline	0.359
Learned ($\lambda = 0.6$)	0.381 (+6.1%)*

Table 2: Baseline and learned λ values for interpolating between Terrier and PageRank scores (expanding 20 related articles). Improvements are statistically significant.

Focusing specifically on PageRank scores (expansion of 20 related articles), we conducted a series of cross-validation experiments to try and automatically learn λ settings. Topics were divided into five folds, stratified in such a way that each fold contains a proportional representation of each template. We trained on four folds and tested on the fifth (selecting the λ that maximized the metric in question); the process was repeated five times. We conducted three separate cross-validation runs, tuning on each metric. Results are shown in Table 2, along with relative improvements over baseline (original Terrier rankings); all improvements are statistically significant ($p < 0.05$). We confirm that it is indeed possible to obtain optimal performance for a particular metric given appropriate training data.

6 Discussion

Our experiments suggest that PageRank is more effective than HITS for analyzing the link structure of related document networks. This makes sense, as the notion of hubs and authorities does not find a natural analog in our application (perhaps the closest is "surveys" and "seminal works"). Whereas HITS assumes a particular linking behavior (which may be true of Web authors), PageRank models a random walk over an arbitrary graph structure—and appears equally applicable to explicit Web hyperlinks as well as automatically-computed content similarity links.

Based on this work, we see a number of future directions worth exploring. Our current approach builds related document networks directly from the ranked list—the result is a link graph that is query-biased, i.e., it represents the local neighborhood around a particular region of the document space. We do not know if this is an essential component of our scoring model, or if alternative formulations are equally effective. One might, for example, perform link analysis over the entire document collection (thus generating scores that are query independent). This is likely the preferred approach for operational environments, as it avoid link analysis on-the-fly (since scores can be cached). Although MEDLINE (currently containing over 17 million records) is relatively small by Web standards, we currently lack the computational resources to perform either PageRank or HITS on the entire document collection.

Another interesting possibility is to use related document networks not only for rescoring, but also for expanding the result set. In our reranking experiments, we simply ignored nodes in the related

document network not contained in the original ranked list. These nodes might also be retrieved, although it is unclear how they might be incorporated into the ranked list since by construction they score low with respect to the user's query.

7 Related Work

The related article search feature in PubMed is an instance of "query by example" and can also be understood as a form of single-point relevance feedback. Many commercial search engines provide similar capabilities, through links labeled "similar pages" or "more like this". A number of studies have demonstrated the effectiveness of this feature as a browsing tool (Wilbur and Coffee, 1994; Smucker and Allan, 2006) using simulations of searcher behavior. However, the focus has been on interactive tools for navigating text collections, and not on result ranking.

Cluster-based retrieval has historically received much attention in the information retrieval literature, most recently in the language modeling framework (Liu and Croft, 2004). Clustering can also be used as an interactive search tool, as in Scatter/Gather (Hearst and Pedersen, 1996); cf. (Leuski, 2001). Despite similar goals (exploiting inter-document relationships), clustering represents a different approach from this work. Clustering algorithms typically function by grouping together similar documents based on a high-dimensional vector representation. Thus, the relationship of interest is group membership (i.e., a document is a member of the group defined by all documents in the cluster). In contrast, related document networks focus on pairwise content similarity, and gives rise to different algorithms for exploiting structure.

This work is most similar to that of Kurland and Lee (2005), who rerank documents based on generation links induced from language models. Diaz (2007) describes the process of score regularization, based on the idea that similar document should have similar retrieval scores. These represent alternative methods for exploiting the link structure of document networks.

In addition to IR applications, link analysis has also been adapted for NLP tasks. For example, LexPageRank (Erkan and Radev, 2004) computes PageRank scores over a network defined by sentence cosine similarity, and has been shown to outperform centroid-based techniques for extractive summarization. Other applications of graph-based algorithms in summarization include (Mihalcea, 2004).

8 Conclusion

We demonstrate that in the absence of explicit hyperlinks, it is possible to exploit networks defined by automatically-generated content-similarity links for document retrieval. Evidence from link structure derived from PageRank scores can be integrated with standard retrieval scores in a straightforward manner, and yields significant improvements in ranked retrieval performance. In the context of PubMed, related document networks represent a simple extension of existing functionality.

9 Acknowledgments

This work was funded in part by the National Library of Medicine. I would like to thank W. John Wilbur and Michael DiCuccio for engaging discussions. This work would not have been possible without the kind support of Esther and Kiri.

References

G. Amati and C. van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389.

F. Diaz. 2007. Regularizing query-based retrieval scores. *Information Retrieval*, 10(6):531–562.

G. Erkan and D. Radev. 2004. LexPageRank: prestige in multi-document text summarization. In *Proceedings of EMNLP 2004*, 365–371.

M. Hearst and J. Pedersen. 1996. Reexaming the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of SIGIR 1996*, 76–84.

W. Hersh, A. Cohen, J. Yang, R. Bhupatiraju, P. Roberts, and M. Hearst. 2005. TREC 2005 Genomics Track overview. In *Proceedings of TREC 2005*.

X. Huang, M. Zhong, and L. Si. 2005. York University at TREC 2005: Genomics track. In *Proceedings of TREC 2005*.

J. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

O. Kurland and L. Lee. 2005. PageRank without hyperlinks: Structural re-ranking using links induced by language models. In *Proceedings of SIGIR 2005*, 306–313.

A. Leuski. 2001. Evaluating document clustering for interactive information retrieval. In *Proceedings of CIKM 2001*, 33–40.

J. Lin and W. Wilbur. 2007. PubMed related articles: A probabilistic topic-based model for content similarity. *BMC Bioinformatics*, 8:423.

J. Lin, M. DiCuccio, V. Grigoryan, and W. Wilbur. 2007. Exploring the effectiveness of related article search in PubMed. Technical Report CS-TR-4877/UMIACS-TR-2007-36/HCIL-2007-10, University of Maryland, College Park, Maryland.

X. Liu and W. Croft. 2004. Cluster-based retrieval using language models. In *Proceedings of SIGIR 2004*, 186–193.

R. Mihalcea. 2004. Graph-based reanking algorithms for sentence extraction, applied to text summarization. In *Proceedings of ACL 2004, Companion Volume*, 170–173.

R. Nallapati. 2004. Discriminative models for information retrieval. In *Proceedings of SIGIR 2004*, 64–71.

L. Page, S. Brin, R. Motwani, and T. Winograd. 1999. The PageRank citation ranking: Bringing order to the Web. Stanford Digital Library Working Paper SIDL-WP-1999-0120, Stanford University.

P. Pirolli and S. Card. 1999. Information foraging. *Psychology Review*, 106(4):643–675.

M. Smucker and J. Allan. 2006. Find-Similar: Similarity browsing as a search tool. In *Proceedings of SIGIR 2006*, 461–468.

C. van Rijsbergen. 1979. *Information Retrieval*. Butterworths, London, the United Kingdom.

E. Voorhees. 1985. The cluster hypothesis revisited. In *Proceedings of SIGIR 1985*, 188–196.

W. Wilbur and L. Coffee. 1994. The effectiveness of document neighboring in search enhancement. *Information Processing and Management*, 30(2):253–266.