Exploiting Site-Level Information to Improve Web Search

Andrei Broder[†], Evgeniy Gabrilovich[†], Vanja Josifovski[†]
George Mavromatis[†], Donald Metzler[‡], and Jane Wang[†]

† Yahoo! Research, 4301 Great America Parkway, Santa Clara, CA 95054
‡ Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292

ABSTRACT

Ranking Web search results has long evolved beyond simple bag-of-words retrieval models. Modern search engines routinely employ machine learning ranking that relies on exogenous relevance signals. Yet the majority of current methods still evaluate each Web page out of context. In this work, we introduce a novel source of relevance information for Web search by evaluating each page *in the context* of its host Web site. For this purpose, we devise two strategies for compactly representing entire Web sites. We formalize our approach by building two indices, a traditional page index and a new site index, where each "document" represents the an entire Web site. At runtime, a query is first executed against both indices, and then the final page score for a given query is produced by combining the scores of the page and its site. Experimental results carried out on a large-scale Web search test collection from a major commercial search engine confirm the proposed approach leads to consistent and significant improvements in retrieval effectiveness.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval

General Terms: Algorithms, Experimentation

Keywords: Web search, corpus structure, textual features

1. INTRODUCTION

Modern Web search engines routinely crawl and analyze tens of billions of Web pages, yet finding highly relevant results for every query remains a major challenge. Conventionally, Web search is performed in two phases. The first phase, usually implemented as a variant of the bag-of-words approach, seeks high recall and retrieves a set of candidate documents, which contain all the query words. The second retrieval phase seeks high precision by reranking the candidate pages using machine learning techniques that rely on a much larger feature set, including numerous sources of exogenous knowledge. Over the recent years, the learning to rank approach has become a standard technique in many IR tasks beyond Web search, yet one thing remained constant—ranking generally evaluates each document in isolation.

We believe this approach to be quite limiting in Web search, as it overlooks the information encoded in the organization of pages on the Web. Admittedly, existing approaches do incorporate some information collected from the network of hyperlinks, namely, by employing link analysis algorithms such as PageRank [2] or HITS [7], and by aggregating anchor text of incoming links to each page [12]. However, even after augmentation with such external evidence, each page is essentially scored by disregarding its context.

We propose an extension to the existing methods so that each page is considered in its natural context, namely, in the context of its host Web site. In particular, we believe this approach allows us to better identify the parts of the page that are truly representative of its content, as well as those "incidental" parts that can be ignored. Conceptually, our method can be viewed as complementary to PageRank-style graph algorithms. Whereas the latter boost individual page scores based on the overall network prominence of the Web site, our method incorporates textual clues that cannot be captured through link analysis alone. The proposed approach can also help overcome anchor text sparsity. While anchor text has been shown to be a strong relevance signal, many pages have no meaningful incoming anchor text. Aggregating the anchor text at the site level allows for cross-use of anchor text for multiple pages of the same site.

Naturally, indiscriminate aggregation of content over entire Web sites might overshadow key nuggets of information in individual pages, thus ruining the ranking. One could envision multiple ways to incorporate site-level information into the ranking process. One way to do so, which is reminiscent of the traditional page-level ranking, is to use the site information to augment the page representation. A notable drawback of this approach, however, is that the page index becomes prohibitively large owing to massive text duplication, as the site text is added to each page from the site. To this end, we adopted an alternative approach where we maintain two indices: a traditional *URL index* (page index), and a separate (and much smaller) *site index.* The latter index is populated with site representations, which succinctly represent the content of the entire site. Each page is scored with respect to both indices, and the resulting scores are combined and used for ranking. This two-index approach provides an efficient and effective way to augment the page ranking process with site information without having to replicate the expansion data for each page in the index.

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.
Copyright 2010 ACM 978-1-4503-0099-5/10/10 ...$10.00.

The contributions of this paper are threefold. First, we propose a novel search paradigm, which combines evidence from a traditional Web page-based index as well as a site-based index. The site index helps improve retrieval effectiveness by providing more contextually relevant information for the pages. Second, we describe two novel approaches for representing the site content. We do so by using the information external to the site (i.e., the incoming anchor text) and internal to the site (i.e., a sample of the site's own pages). Finally, we perform an empirical evaluation of the proposed approach using a large test collection from a major commercial Web search engine. Experimental results show that our approach leads to significant improvements in retrieval effectiveness.

2. RELATED WORK

There are two main lines of research most closely related to our work. The first body of research has investigated ways of imposing *implicit* structure on corpora that are not explicitly structured (e.g., news corpora). Most of the proposed approaches rely on clustering to define the implicit structure. One of the first approaches, by Jardine and van Rijsbergen [5], clustered documents using agglomerative clustering. The more recent approaches developed by Liu and Croft [9] and Kurland and Lee [8] cluster documents using K-means clustering. After the entire corpus has been clustered, document term weights are computed using document, cluster, and corpus statistics. In these approaches, term weights reflect the importance of the term in the context of the document itself, the context of similar documents, or the corpus as a whole. Other strategies for imposing implicit corpus structure to improve search relevance have been proposed as well, such as using topic modeling [15]. Although using implicit structure for textual matching has been shown to be useful by various researchers, it is infeasible to apply such methods to large collections, such as the Web, primarily due to the prohibitive cost of clustering billions of documents.

The other line of research looked at ways for improving search relevance using *explicit* corpus structure. As we mentioned before, not all document collections are structured, but for those that are, there are certain benefits to using this explicit structure. The key benefit of such approaches is that clustering is not necessary, and that explicit corpus structure is likely to be more accurate and more useful than implicitly defined structure. Most of the research in this direction used the link structure of the Web to develop improved textual matching strategies. Qin et al. [13], Shakery and Zhai [14], and Metzler et al. [12] proposed methods for propagating text matching scores and textual representation across the Web graph. However, such methods suffer from limited coverage due to properties of the Web graph. For example, the anchor text aggregation method proposed by Metzler et al. only covers about 5% of all URLs [12], whereas our proposed approach has the ability to cover 100% of URLs.

Finally, in work that is perhaps the most closely related to ours, Aguiar [1] proposed building two indices, one for pages and another for the context of the pages, where the context is defined using text and link similarity. The latter work is similar to our proposed method in its use of the page context; however, our approaches differ in the way this context is defined. Specifically, we build a site index, which is conceptually more straightforward and less computationally demanding to build and manage than the contextual index proposed by Aguiar. Furthermore, Aguiar's use of text and link similarity is reminiscent of discovering imprecise implicit structure. In contrast, our method uses Web sites as the most natural context of Web pages.

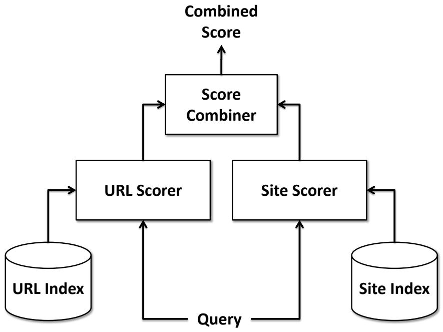

Figure 1: Overview of site-level contextual scoring framework.

3. USING SITE-LEVEL INFORMATION TO IMPROVE SEARCH

This section describes our proposed methodology for using site-level information to improve Web search relevance. We are primarily interested in the textual aspects of Web search relevance, and therefore our ultimate goal is to improve the textual scoring component(s) of a search engine.

Figure 1 provides a high level overview of our framework, which has two primary components. The *indexing component* is responsible for constructing the two search indices, namely, a *URL index* and a *site index*. The URL index is a standard Web search index, where the indexing unit is a Web page. The site index is a novel aspect of our framework. Its indexing unit represents a Web site, as opposed to a Web page. The site index is used to encoded contextual information for all of the Web pages within the site.

The *scoring component* is responsible for executing queries against the URL and site indices. Queries can be executed against the two indices in parallel, reducing the overall latency. The outputs of the two indices are then aggregated to produce a site-specific retrieval score, which can be used directly, or as a feature in a subsequent reranking step.

The remainder of this section will describe how the site index is constructed, as well as how the site-specific retrieval score is computed.

3.1 Site Index

The primary use of the site index is to encode contextual information about the pages within a given Web site. Our hypothesis is that such contextual information will be useful for improving search relevance. Before building a site index, however, we must first determine how to effectively represent an entire Web site. Web sites are diverse, in terms of size, popularity, and topicality. A good site representation will be concise (require minimal storage), topically focused (relevant to the site), and exhaustive (cover all site topics).

The most naïve way to represent a site is to concatenate together the HTML of all of the pages within the site. However, this would result in very large "documents", and the resulting site index would be approximately the same size as the URL index. As we show below, it is possible to build useful site indices that consume only a fraction of the space of the URL index.

One simple strategy for significantly reducing the storage requirements of the site index is to sample URLs from the site and only use the sampled URLs when constructing the site index. This way, sites with hundreds of thousands or even millions of pages will not cause the site index size to explode. We employ this strategy in our experiments. Given a site, say, `www.site.com`, we issue the query "site:www.site.com" to a Web search engine and collect the top 1000 returned URLs, which we use as a sample of the site. Sampling in this way returns URLs that are authoritative and popular pages within the site, thereby serving as a good representation of the site. Of course, other more sophisticated sampling strategies are possible, but since our focus is on the general site-specific scoring framework, we do not explore other strategies here.

We will now describe two different ways to build site representations. It should be noted that we assume that the site index is built using the same indexing software that is used to build the URL index. This way, no specialized indexing software is necessary to implement our proposed approaches.

The first representation that we consider is called the *page-based site representation*. Given a site, the HTML of all of the pages from the site, or a sample of pages from the site, is concatenated together. As we mentioned before, this is perhaps the simplest way to represent a site, but certainly not the most concise or useful, as the representation will contain many extraneous terms that are not characteristic of the site (e.g., legal disclaimer terms repeated on every page) and hence do not provide useful contextual evidence.

Our second representation, which we refer to as the *anchor text-based site representation*, attempts to overcome some of the problems associated with the page-based index. In this representation, the external anchor text for all the pages from the site (or for a sample of pages) is concatenated together. This site representation is expected to be much more concise as well as precise, since anchor text tends to be very focused.

Another reason for proposing this representation is that anchor text is known to be one of the most important textual sources of evidence for Web search, given its lexical similarity to queries. Recently, Metzler et al. [12] posed the anchor text sparsity problem, which highlights the fact that most Web pages have very little, if any, anchor text, and that retrieval effectiveness can be significantly improved by adding anchor text aggregated across the Web graph to Web page representations. Thus, our anchor text-based site index is an alternative approach for associating more relevant anchor text with Web pages, and therefore can be thought of as another strategy for overcoming anchor text sparsity.

3.2 Site Specific Scoring and Features

The other major constituent of our proposed framework is the scoring component. This component is responsible for computing retrieval scores using the URL and site indices, and then combining the two individual scores together. Since we are primarily interested in formulating a general framework for using site information, here we propose a simple combination mechanism as a proof of concept, which can easily be built upon in future work. Our combination mechanism can be formally described as follows:

$$f(Q,U) = (1-\lambda)\cdot S_{url}(Q,U) + \lambda \cdot S_{site}(Q, site(U)) \quad (1)$$

where Q is the user's query, U is the page being scored, $S_{url}(Q,U)$ is the URL index score, $S_{site}(Q, site(U))$ is the score based on the site index, and λ is a free parameter that controls the combination. Thus, the combined score is simply a linear mixture of the URL and site index scores.

The score $f(Q,U)$ can either be used directly to rank documents or used as a feature within a machine learned ranking function.

The ranking function that we utilize to compute $S_{url}(Q,U)$ and $S_{site}(Q, site(U))$ is called BM25F-SD. It is a novel combination of BM25F [16] weighting and Metzler's sequential dependence (SD) model [11], which provides an effective framework for term proximity matching. BM25F-SD assigns different weights to matches due to different document fields (e.g., title, body, anchor text, etc.) and also boosts phrase and proximity matches that occur within each field. Both of these factors have been shown to be important for effective textual matching in Web search. The BM25F-SD scoring function is defined as:

$$\begin{aligned} S(Q,U) &= \lambda_T \sum_{w\in Q} wt(w,U) + \\ &\quad \lambda_O \sum_{w_i,w_{i+1}\in Q} wt\text{``}w_i w_{i+1}\text{''}, U) + \\ &\quad \lambda_U \sum_{w_i,w_{i+1}\in Q} wt(prox(w_i,w_{i+1}),U) \quad (2) \end{aligned}$$

where $wt(w,U)$ is the BM25F weight of the term w in page U, $wt(\text{``}w_i w_{i+1}\text{''}, U)$ is the BM25F weight of the exact phrase "$w_i w_{i+1}$" in page U, and $wt(prox(w_i, w_{i+1}), U)$ is the BM25F weight of terms w_i and w_{i+1} occurring within a window of 8 terms of each other (this is the proximity component). Furthermore, λ_T, λ_O, and λ_U are free parameters that control the influence of each type of match on the scoring. Additional details can be found in [10].

4. EMPIRICAL EVALUATION

We now empirically evaluate our proposed approach using a large, real-world Web search test collection from a commercial search engine.

4.1 Data Description and Metrics

Our test collection is divided into a training and test set. The training set consists of 20,120 queries and 416,183 query-URL pairs, while the test set is made up of 3,556 queries and 139,940 query-URL pairs. The queries were randomly sampled from the query log of a major commercial search engine. For each query-URL pair, human relevance judgments were obtained using a five point scale (Perfect, Excellent, Good, Fair, and Bad).

All experiments make use of the same URL index, which is a standard Web search index. Documents within the URL index are treated as structured documents. That is, each document consists of a collection of fields, including title, body, headings, etc. Furthermore, all URLs are accompanied with their (external) anchor text, if available.

Site indices were built using the two site representations formulated in Section 3.1. Each index consists of the sites found in the union of the training and test sets. This results in a total of 207,222 sites.

We evaluate our proposed site-specific ranking methods using the human judgments associated with the query-URL

Metric	URL	Page	Anchor
DCG@1	3.8947	3.9528^u	3.9720^u
DCG@5	9.1165	9.1956^u	9.2510^{up}
DCG@10	12.3378	12.4047^u	12.4757^{up}
DCG	20.9716	21.0261^u	21.0661^{up}
NDCG@1	0.5270	0.5362^u	0.5395^u
NDCG@5	0.5185	0.5228^u	0.5272^{up}
NDCG@10	0.5543	0.5576^u	0.5607^{up}
NDCG	0.7435	0.7456^u	0.7472^{up}
ERR@1	0.3064	0.3107^u	0.3124^u
ERR@5	0.4136	0.4172^u	0.4193^{up}
ERR@10	0.4325	0.4359^u	0.4378^{up}
ERR	0.4415	0.4447^u	0.4465^{up}

Table 1: Summary of BM25F-SD ranking function results. The superscripts u and p denote statistically significant improvements over the URL only index and page-based site index, respectively.

pairs in the test set. To provide a comprehensive view of the results, we evaluate our methods using DCG, NDCG [6], and ERR [3], which are commonly used to evaluate Web search engines. The gains used in the DCG and NDCG computation for URLs judged Perfect, Excellent, Good, Fair, and Bad are 10, 7, 3, 0.5, and 0, respectively. Following [3], the judgments were mapped to the following probabilities for use with ERR: 0.9375, 0.4375, 0.1875, 0.0625, and 0. Statistical significance was tested using a paired, one-tailed non-parametric bootstrap test [4]. All tests were done at the $p < 0.05$ level.

4.2 Experimental Results

The results of our experiments using the BM25F-SD ranking function are shown in Table 1. The URL column represents the baseline system, which only uses the URL index for retrieval (i.e., $\lambda = 0$ in Equation 1). The Page and Anchor columns correspond to the case where the page-based site index and anchor text-based site index, respectively, are used as an additional source of evidence for ranking (i.e., $\lambda > 0$ in Equation 1).

There are several conclusions that can be drawn from these results. First, and most importantly, is that all approaches that used information from site indices showed significant improvements across all metrics compared to the baseline. Therefore, there is a clear, consistent, and significant benefit to using site-level information on top of a sophisticated text-only ranking function.

Additionally, these results suggest that the anchor text-based site representation is far superior to the page-based site index. The improvements over the baseline in terms of NDCG@1, NDCG@5, NDCG@10, and NDCG are 2.4%, 1.7%, 1.2%, and 0.5%, respectively. Similarly, the observed improvements for ERR@1, ERR@5, ERR@10, and ERR are 2.0%, 1.4%, 1.2%, and 1.2%, respectively. Using the anchor text-based site index resulted in statistically significant improvements compared to the page-based site index for 9 out of the 12 metrics.

Analagous experiments were carried out using a language modeling ranking function and a machine learned ranking function. Due to space limitations, we only provide a brief summary of the results. Statistically significant improvements were observed for both ranking functions, suggesting that our proposed approach can be useful in a variety of search settings.

5. CONCLUSIONS AND FUTURE WORK

We presented a novel methodology for improving Web search relevance using site-level information. Our approach uses a site index in addition to a traditional URL index for Web search ranking. We described two strategies for representing and indexing Web sites. We also evaluated the effectiveness of combining evidence from the site index with the URL index using a sophisticated text-only ranking function. Our experiments were carried out over a very large Web search test collection from a major commercial search engine. The results showed that using site-level information can consistently and significantly improve Web search effectiveness.

As part of future work, we are interested in exploring different strategies for sampling URLs from sites beyond the simple approach used in this study. We would also like to investigate additional site-level features that could be computed using the site index. Finally, we believe it is important to develop a better understanding of the role of site cohesiveness within our framework, especially for the anchor text site index.

6. REFERENCES

[1] F. Aguiar. Improving web search by the identification of contextual information. In *Studies In Fuzziness And Soft Computing*, Studies In Fuzziness And Soft Computing. Physica-Verlag Heidelberg, 2003.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[3] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. 18th Intl. Conf. on Information and Knowledge Management*, page To appear., 2009.

[4] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, 1993.

[5] N. Jardine and C. van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7(5), 1971.

[6] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

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

[8] O. Kurland and L. Lee. Corpus structure, language models, and ad hoc information retrieval. In *SIGIR*, pages 194–201, 2004.

[9] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *SIGIR*, pages 186–193, 2004.

[10] D. Metzler. *Beyond Bags of Words: Effectively Modeling Dependence and Features in Information Retrieval.* PhD thesis, University of Massachusetts, Amherst, MA, 2007.

[11] D. Metzler and W. B. Croft. A Markov Random Field model for term dependencies. In *Proc. 28th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 472–479, 2005.

[12] D. Metzler, J. Novak, H. Cui, and S. Reddy. Building enriched document representations using aggregated anchor text. In *Proc. 32nd Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 219–226, New York, NY, USA, 2009. ACM.

[13] T. Qin, T.-Y. Liu, X.-D. Zhang, Z. Chen, and W.-Y. Ma. A study of relevance propagation for web search. In *SIGIR*, pages 408–415, 2005.

[14] A. Shakery and C. Zhai. Smoothing document language models with probabilistic term count propagation. *Information Retrieval*, 11(2), 2008.

[15] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *SIGIR*, pages 178–185, 2006.

[16] H. Zaragoza, N. Craswell, M. Taylor, S. Saria, and S. Robertson. Microsoft Cambridge at TREC 13: Web and hard tracks. In *Proc. 13th Text REtrieval Conference*, 2004.