Query-Independent Evidence in Home Page Finding

TRYSTAN UPSTILL
Australian National University
and
NICK CRASWELL and DAVID HAWKING
CSIRO Mathematical and Information Sciences

Hyperlink recommendation evidence, that is, evidence based on the structure of a web's link graph, is widely exploited by commercial Web search systems. However there is little published work to support its popularity. Another form of query-independent evidence, URL-type, has been shown to be beneficial on a home page finding task. We compared the usefulness of these types of evidence on the home page finding task, combined with both content and anchor text baselines. Our experiments made use of five query sets spanning three corpora—one enterprise crawl, and the WT10g and VLC2 Web test collections.

We found that, in optimal conditions, all of the query-independent methods studied (in-degree, URL-type, and two variants of PageRank) offered a better than random improvement on a content-only baseline. However, only URL-type offered a better than random improvement on an anchor text baseline. In realistic settings, for either baseline, only URL-type offered consistent gains. In combination with URL-type the anchor text baseline was more useful for finding popular home pages, but URL-type with content was more useful for finding randomly selected home pages. We conclude that a general home page finding system should combine evidence from document content, anchor text, and URL-type classification.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*

General Terms: Experimentation

Additional Key Words and Phrases: Web information retrieval, citation and link analysis, connectivity

1. INTRODUCTION

Analysis of a sample of search requests submitted to the search engines of a university and a large media corporation showed that nearly 60% of the former and 29% of the latter apparently represented attempts to name an

Authors' addresses: T. Upstill: Department of Computer Science, The Australian National University, Canberra, ACT, Australia 0200; email: trystan@cs.anu.edu.au; N. Craswell and D. Hawking: CSIRO Mathematical and Information Sciences, GPO Box 664, Canberra, ACT, Australia, 2601; email: {Nick.Craswell,David.Hawking}@csiro.au.

entity.[1] These queries typically specified entities such as people, companies, departments, and products (e.g. "Trystan Upstill," "CSIRO," "Computer Science," or "Panoptic").

A searcher who submits an entity name as a query is very likely to be pleased to find a home page for that entity at the top of the list of search results. Home pages provide primary-source information in response to *informational* queries and are the only correct answers for *navigational* queries [Travis and Broder 2001; Broder 2002].

This paper addresses the home page finding problem. It assumes that incoming queries are attempts to navigate to the home page of a particular site. This might occur in practice if the search interface allowed searchers to specify home page search, for example by ticking a box. Home page finding techniques studied here could also be combined with other ranking methods in a general search system.

Some potentially useful evidence for home page finding is query-dependent. This includes the presence of query words in the document's text, in its referring anchor text (the words you click on in your browser to follow a hyperlink), or in the document's URL (its Web address, e.g., `www.mapquest.com`). It is known that full text relevance ranking is not particularly effective for home page finding [Craswell et al. 2001; Hawking and Craswell 2001; Singhal and Kaszkiel 2001].

Other potentially useful evidence is query-independent. This was demonstrated in the TREC-2001 home page finding task [Hawking and Craswell 2001]. The best run was submitted by UTwente/TNO [Westerveld et al. 2001] and used a page's URL-type as evidence. The authors computed the probability of a page being a home page given the type of its URL: root, subroot, path, or file. (See Section 2.1.) They then combined this query-independent probability with a query-dependent (content or anchor text) score.

Hyperlink recommendations (scores computed from the structure of a link graph) are also query-independent (e.g., In-degree and PageRank [Page et al. 1998; Brin and Page 1998]). This type of evidence is generally perceived to be a method by which the "quality" of query results is improved [Brin and Page 1998], and is widely used in commercial Web search systems such as Google (`www.google.com`) and FAST (`www.alltheweb.com`). These search engines have been shown to perform well on a home page finding task [Hawking et al. 2001] (although these facts may not be related). Furthermore, link evidence also tends to correlate with the URL-type evidence used by UTwente/TNO, as shown in Figure 1. However, there exists little empirical evidence that hyperlink recommendation improves the overall quality of search results[2] [Amento et al. 2000]. TREC participants were unable to demonstrate improvements through use of hyperlink evidence on traditional TREC relevance tasks [Hawking et al. 1999].

[1]These proportions are considerably higher than the corresponding percentage (around 15%) of a sample of queries submitted to the Internet search engine `www.alltheweb.com` [Fast Search and Transfer ASA 2002]. Note that the proportion is difficult to quantify precisely because of language mixture, typographical errors, and proper names which are indistinguishable from ordinary words.
[2]Note that Page et al. [1998] only evaluated PageRank by presenting some example search results.

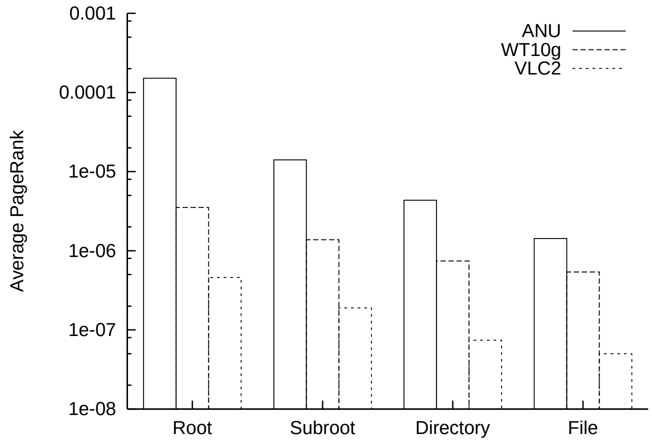

Fig. 1. Average Democratic PageRank (DPR) for each of the URL classes identified by Westerveld et al. [2001] over the ANU, WT10g and, VLC2 corpora. These authors found that Root pages had the highest probability of being home pages and File pages had the least. PageRank scores shown here generally reflect this.

The scope of our experiments is as follows:

—five home page finding test collections based on three crawls ranging between 400,000 and 18 million pages;
—four types of query-independent evidence: URL-type, in-degree, Democratic PageRank, and Aristocratic PageRank;
—two query-dependent baselines: content and anchor text; and
—three methods for combining query-dependent and query-independent evidence; two realistic combinations and the optimal combination.

The contributions within this scope are as follows:

(1) Experiments to determine which query-independent evidence is most useful in home page finding. This involves an investigation of both the maximum possible improvement and the realistically achievable improvement offered by each type of query-independent evidence over a wide range of datasets.
(2) Analysis to determine which methods are useful for finding well-known home pages, and which are useful for finding randomly sampled home pages.
(3) Analysis of the relationship between in-degree and PageRank.
(4) Evaluation of link methods on a small crawl using larger-crawl link information.

2. QUERY-INDEPENDENT EVIDENCE

The focus of the present study is on evaluating the potential contribution to home page finding of query-independent evidence: URL-type, in-degree, and PageRank.

2.1 URL-Type

We followed UTwente/TNO and classified URLs (after stripping off a trailing `index.html`, if present) into four categories:

—*root*. a domain name, e.g. `www.cyborg.com/`;
—*subroot*. a domain name followed by a single directory, for example, `www.glasgow.ac.uk/staff/`;
—*path*. a domain name followed by two or more directories, for example, `trec.nist.gov/pubs/trec9/`;
—*file*. any URL ending in a filename rather than a directory, for example, `trec.nist.gov/contact.html`.

There are therefore only four possible values for this URL-type evidence. We used the same ranking of values as did TNO/UTwente, that is, root > subroot > path > file.

2.2 URL Canonicalization

The success of link-based methods, whether query-dependent or not, can depend critically on the robustness of the methods employed to *canonicalize* URLs. This is the process by which different but equivalent URLs specified as hyperlink targets are converted to a canonical form. For example: `sony.com`, `http://www.sony.com/`, `http://www.sony.com:80/`, and `http://www.sony.com:80/index.html` might all be represented as `www.sony.com/`. Failing to recognize such equivalences can lead to phantom structures within the link graph and to incorrect assignment of both anchor text and hyperlink-derived scores. Ideally all URL equivalences due to duplications and redirects would be taken into account; however the necessary information was not recorded during the gathering of the corpora and was therefore unavailable for use.

2.3 In-Degree

A page's in-degree is computed simply by counting its incoming links [Carrière and Kazman 1997]. When working within a subset of the Web, observed in-degrees may dramatically underestimate the full-Web values. This is because some link targets may be popular outside the immediate Web community.

Westerveld et al. [2001] have previously investigated the use of page in-degree (the number of links referring to a page) for the home page finding task, but found it to be less useful than URL-type evidence. We include it here for comparison with PageRank.

2.4 PageRank

PageRank is a more sophisticated query-independent link citation measure developed by Page and Brin [Page et al. 1998; Brin and Page 1998] to "objectively and mechanically [measure] the human interest and attention devoted [to Web pages]" [Page et al. 1998, p. 1]. PageRank is believed to be the primary link recommendation scheme employed in the Google search engine and search appliance.

PageRank simulates the behavior of a "random Web surfer" [Page et al. 1998] who navigates by randomly following links. If a page with no outgoing links is reached, the surfer jumps to a randomly chosen bookmark. In addition to this normal surfing behavior, the surfer occasionally spontaneously jumps to a bookmark instead of following a link. The PageRank of a page is the probability that the Web surfer will be visiting that page at any given moment. A formal description of the PageRank algorithm is [Brin and Page 1998; Page et al. 1998]:

$$
\begin{array}{l}
\vec{R}_0 \leftarrow \vec{S} \\
\mathit{loop}: \\
\quad r \leftarrow \mathit{dang}(\vec{R}_i) \\
\quad \vec{R}_{i+1} \leftarrow r\vec{E} + A\vec{R}_i \\
\quad \vec{R}_{i+1} \leftarrow (1-d)\vec{E} + d(\vec{R}_{i+1}) \\
\quad \delta \leftarrow \|\vec{R}_{i+1} - \vec{R}_i\|_1 \\
\mathit{while}\ \delta > \epsilon
\end{array}
$$

Where $\vec{R}_i$ is the PageRank vector at iteration i, A is the link adjacency matrix, $\vec{S}$ is the initial PageRank vector, $\vec{E}$ is the vector of bookmarked pages, *dang*() is a function that returns the PageRank of all nodes that have no outgoing links, r is the amount of PageRank lost due to dangling links which is distributed among bookmarks (after Ng et al. [2001]), d is a constant which controls the proportion of random noise (spontaneous jumping) introduced into the system to ensure stability ($0 < d < 1$), and ϵ is the convergence constant. In our formulation bookmarks receive the $(1-d)$ random noise on each iteration, thereby maximizing the effect of bookmarks ($\vec{E}$). The double bar ($\|_1$) notation indicates an l_1 norm, the sum of a vector element's absolute values.

2.4.1 *PageRank Variations: Democratic and Aristocratic.* For a given link graph, PageRank varies according to the values of the d constant and the set of bookmark pages $\vec{E}$. In our implementation we set $d = 0.85$. We implemented two different PageRank schemes described by Brin and Page by varying the bookmark vector ($\vec{E}$). The first variation is a "democratic" unbiased PageRank in which all pages are *a priori* considered equal. The second is an "aristocratic" PageRank in which the PageRank calculation is customized using bookmark pages from a hand-picked source.

In Democratic PageRank, or DPR, every page in the collection is considered to be a bookmark and every page has a nonzero PageRank. Similarly, every link is important and thus in-degree is a good predictor of DPR. Because it is easy for Web page authors to create links, it is easy to manipulate DPR with link spam.[3]

In Aristocratic PageRank, or APR, a set of authoritative pages are used as bookmarks to systematically *bias* scores. In practice the authoritative

[3]*spam* is the name applied to techniques used by Web publishers to artificially boost the rank of their pages. Such techniques include addition of otherwise unneeded keywords and hyperlinks.

pages might be taken from a reputable directory service, such as Yahoo! (`www.yahoo.com`) [Yahoo! n.d.], Looksmart (`www.looksmart.com`) [Looksmart n.d.], or the Open Directory (`dmoz.org`) [DMOZ n.d.]. This would tend to give such pages higher APR. Further, APR may be harder to spam because newly created pages are not included in the bookmarks by default.

2.4.2 *How Large a Link Graph is Needed for PageRank?.* PageRanks can be calculated for a web graph of any size. PageRank scores are therefore usable within any web crawl,[4] including single organizations (enterprise) and portals. The recently released Google organizational search appliance incorporates PageRank for crawls of below 150,000 pages [Google 2002a].

It is sometimes claimed that PageRanks are not useful unless the link graph is very large (tens or hundreds of millions of nodes) but this has not been documented. We investigated this question using three different test collection sizes (and also by measuring the correlation between locally computed PageRanks and global Web PageRanks reported by Google). While for practical reasons we cannot test on a billion page crawl, the overwhelming majority of crawls used by operational search systems are no larger than those investigated here.

3. EXPERIMENTAL FRAMEWORK

We now describe the home page finding retrieval task, the baselines and measures used to evaluate retrieval effectiveness, and the test collections used in our experiments and their salient properties. Following this we outline the ways in which query-independent and query-dependent information are combined.

3.1 Retrieval Task and Measures

The home page finding task is as defined in the TREC-2001 Web Track [Hawking and Craswell 2001] and in Craswell et al. [2001]. An example of a home page finding search is when a user wants to visit `trec.nist.gov` and types the query `Text Retrieval Conference`. The task is similar to Bharat and Mihaila's organization search [Bharat and Mihaila 2001], where users provided Web site naming queries, and Singhal and Kaszkiel's site finding experiment [Singhal and Kaszkiel 2001], where queries were taken from an Excite log [Excite 2002].

Page and Brin's *common case* queries [Page et al. 1998] are related, in that home pages are highly valued answers, but differ from the present task in that the query is a generic name or prescription rather than the name of a specific entity. An example of a common case query might be `flowers` (cf. `Interflora`) to which a good response would be a selection of popular commercial florist's home pages.

In our experiments, we measured success rates at several cutoffs. The success rate measure is indicated by S@n where n is the cutoff rank. S@10 measures how often the correct page was returned within the first 10 results, and S@1 corresponds to the probability that the right answer appeared at rank 1 (cf. the "I'm feeling lucky" button on Google). We also use S@5, which represents

[4] A crawl is the set of pages.

how often the correct answer might be visible in the first results page without scrolling ("above the fold").

We performed a Wilcoxon matched-pairs signed ranks test to determine whether improvements afforded by a reranking were significant. This test compares the algorithms by the differences in ranks achieved rather than by their success rates. Throughout our experiments we used a confidence level of 95% ($\alpha = 0.05$).

3.2 Baselines

The first query-dependent baseline was an Okapi BM25 ranking of document content, termed *content*. The second was an Okapi BM25 ranking of surrogate documents each consisting of aggregated anchor text descriptions (from all pages pointing to that document) of a page, termed *anchor text*. Anchor text comprises the words that a user clicks on in order to follow a link, but none of the text contained within the target page. To reduce the complexity of our anchor text experiment, we did not consider the text surrounding a link (following from Craswell et al. [2001]).

The Okapi BM25 relevance scoring formula (see Appendix A.2) is due to Robertson et al. [1994, pages 110–111] and has proven consistently effective in TREC evaluations. It takes into account the number of times a query word occurs in a document, the proportion of other documents which also contain the query word, and the relative length of the document. Feedback and stemming were not employed (to maintain consistency with previous home page finding experiments [Craswell et al. 2001]).

3.3 Controls and Experimental Conditions

Each experimental condition constituted a reranking of the top section of a baseline ranking of documents on the basis of a query-independent variable.

The experimental conditions are labeled as follows:

—*Indeg*. a re-ranking by in-degree;
—*DPR*. a re-ranking by Democratic PageRank;
—*APR*. a re-ranking by Aristocratic PageRank with bookmarks from Yahoo! or other directory listings, as would be available for a real system incorporating PageRank;
—*URL*. a re-ranking by the UTwente/TNO URL-type. (root > subroot > path > file).

Note that if scores are equal on the reranking measure, the original baseline ordering is preserved.

3.4 Test Collections

The test corpora used in our evaluation included a recent crawl of a university, plus the VLC2 [Hawking 2000] and WT10g [Bailey et al. 2003] test collections used in the TREC Web track. Detailed collection information is reported in Table I.

Table I. Collection Information (We submitted two sets of queries to the VLC2 collection—a *popular* set (VLC2P) and a *random* set (VLC2R) (see text for explanation). The two sets submitted to WT10g were the set used by Craswell et al. [2001] (WT10gC) and the official queries used in the TREC-2001 home page task (WT10gT). The values in the Content and Anchor queries columns report the number of home pages found by the baseline out of the number of queries submitted (this is equivalent to S@1000 as we only consider the top 1000 results for each search).)

Collection	Size	Pages (million)	Links (million)	Dead links	Content queries	Anchor queries	No. of Bookmarks (APR)
ANU	4.7GB	0.40	6.92	0.646	97/100	99/100	439
WT10gC	10GB	1.69	8.06	0.306	93/100	84/100	25487
WT10gT	10GB	1.69	8.06	0.306	136/145	119/145	25487
VLC2P	100GB	18.57	96.37	3.343	95/100	93/100	77150
VLC2R	100GB	18.57	96.37	3.343	88/100	77/100	77150

Although there are many spam pages on the Web, we found little spam in the three corpora. Any spam-like effect we observed seemed unintentional. For example the pages of a large bibliographic database all linked to the same page, thereby artificially inflating its in-degree and PageRank.

In each run, sets of 100 or more queries were processed over the applicable corpus using the chosen baseline algorithm, and the first 1000 results for each were recorded. While all queries have only one correct answer, that answer may have multiple correct URLs, for example, a host with two aliases. If multiple correct URLs were retrieved, we used the maximum baseline rank and assigned to it the best query-independent score of all the equivalent URLs. This approach introduced a slight bias in favor of the reranking algorithms, to ensure that any beneficial effect would be detected.

We evaluated two important types of scenarios in home page finding, queries for *popular* and *random* home pages. *Popular* queries allowed us to study which forms of evidence allowed effective ranking for queries targeting higher profile sites. *Random* queries allowed us to study effective ranking for any home page, even if it was not well known.[5]

The ANU collection is a deep crawl of a university web in which links external to the university were not followed. The ANU web includes a number of official directories of internal sites, which can be used as bookmark files. This allowed us to observe the behavior of APR in a single-organization environment. Test home pages were picked randomly from these directories and then queries were generated by hand. Consequently we would expect performance of APR to be very good on this collection.

The WT10g corpus is a highly connected collection with a high density of interserver links. The query set labeled WT10gC was created by Craswell et al. [2001] by randomly selecting pages within the corpus, navigating to the corresponding home page, and formulating a query based on the home page's name. The WT10gC set was used as training data in the TREC-2001 Web Track.

[5]Note that the labels *popular* and *random* were chosen for simplicity and were derived from the method used to choose the target answer, not from the nature of the queries. Information about query volumes was obviously unavailable for the test collections and was not used in the case of ANU.

The query set labeled WT10gT was developed by the NIST assessors for the TREC-2001 Web Track using the same method. Westerveld et al. [2001] have previously found that the URL method improved retrieval performance on the WT10gT queries. In our experiments every Yahoo-listed page in the WT10g collection is bookmarked in the APR calculation. These are lower-quality bookmarks than the ANU set as the bookmarks played no part in the selection of either query set. WT10g is a subset of VLC2, and this allows us to observe how the hyperlink methods behave as collection size varies.

The 100gB VLC2 corpus contains roughly one-third of the Internet Archive crawl from February 1997. We evaluated two sets of queries over the VLC2 collection, *popular* (VLC2P), and *random* (VLC2R). The *popular* series was derived from the Yahoo! directory. The *random* series was selected using the method described above for WT10g. For the APR calculation, every Yahoo-listed page in the collection was bookmarked. As such, they are well matched to the VLC2P queries (also from Yahoo!), but less so for VLC2R.

The ANU and VLC2P home pages are considered *popular* because they are derived from directory listings. Directory listings have been chosen by a human editor as important, possibly because they are pages of interest to many people. Such pages also tend to have above average in-degree. This means that more Web page editors have chosen to link to these pages, directing Web surfers (and search engine crawlers) to them.

On all these collections, anchor text ranking has been shown to improve home page finding effectiveness (relative to content-only) [Craswell et al. 2001; Bailey et al. 2003].

3.5 Reranking Baseline Query Results Using Query-Independent Features

Evaluating the usefulness of query-independent evidence in boosting search effectiveness is complicated by the need to combine the query-independent score with a query-dependent score. There is a risk that a spurious negative conclusion could result from a poor choice of combining function.

Accordingly, we gauged the maximum improvement possible due to the query-independent evidence by locating the right answer in the baseline (obviously not possible in a practical system) and reranking it and the documents above it on the basis of the query-independent score alone (*Optimal combination*). No linear combination or product of query-independent and query-dependent scores (assuming positive coefficients) could improve upon this. This is because documents above the right answer score as well or better on both query-independent and query-dependent components (see Figure 2). In Optimal rerankings, a control condition *Random* was introduced in which the correct document and all those above it were arbitrarily shuffled.

We also considered a more realistic scheme in which documents were reranked above a fixed cutoff. The cutoff was trained on one collection and evaluated on others (*Realistic reranking*).

4. OPTIMAL COMBINATION EXPERIMENTS

Figure 2 illustrates the Optimal combination reranking process. This scheme is unrealistic because the resorting relies on knowing the position of the correct

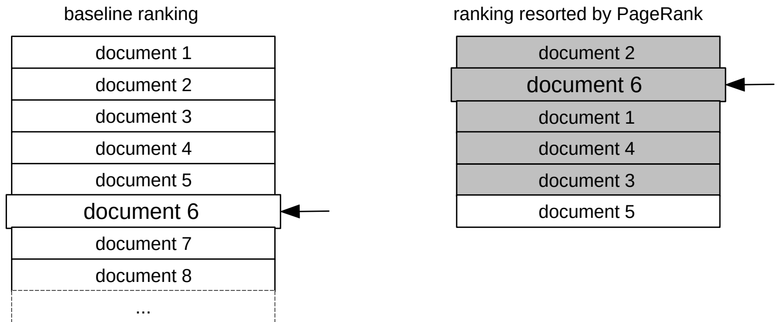

Fig. 2. Example of Optimal reranking. In the baseline, the correct answer is document 6 at rank 6. Reranking by PageRank puts it at rank 2. This is optimal because any document ranked more highly must score as well or better on both baseline and PageRank. In this case, S@5 fails on the baseline and succeeds on reranking. However, a random resorting of the top six would have succeeded in five of six cases, so S@5 for the Random control is (on average) 5/6.

Table II. Optimal Cutoff Results (The Optimal combination experiment is described in Section 4. The best combinations are highlighted in bold. An asterisk signifies that the improvement relative to the random control was significant.)

		Content						Anchor text					
Coll.	Meas.	Base	Rand	Indeg	DPR	APR	URL	Base	Rand	Indeg	DPR	APR	URL
ANU	S@1	0.28	0.37	0.72*	0.65*	0.75*	0.67*	0.76	0.85	0.88	0.89	**0.91**	0.86
ANU	S@5	0.50	0.61	0.88*	0.90*	0.91*	0.87*	0.96	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
ANU	S@10	0.58	0.69	0.93*	0.94*	0.96*	0.91*	0.98	0.98	0.98	0.98	**0.99**	0.98
WT10gC	S@1	0.23	0.34	0.61*	0.59*	0.54*	**0.76***	0.48	0.58	0.62	0.61	0.63	0.73*
WT10gC	S@5	0.45	0.58	0.86*	0.82*	0.84*	**0.89***	0.69	0.73	0.72	0.72	0.73	0.81*
WT10gC	S@10	0.55	0.68	0.86*	0.87*	0.87*	**0.93***	0.72	0.76	0.74	0.75	0.75	0.83*
WT10gT	S@1	0.22	0.34	0.64*	0.62*	0.54*	**0.79***	0.54	0.60	0.63	0.62	0.64	0.72*
WT10gT	S@5	0.48	0.61	0.81*	0.83*	0.80*	**0.88***	0.68	0.73	0.72	0.71	0.75	0.78*
WT10gT	S@10	0.59	0.69	0.86*	0.87*	0.84*	**0.92***	0.72	0.76	0.76	0.77	0.76	0.79*
VLC2P	S@1	0.27	0.38	0.65*	0.61*	0.65*	0.70*	0.69	0.76	0.77	0.78	**0.84**	0.80
VLC2P	S@5	0.51	0.65	0.78*	0.79*	0.80*	0.87*	0.85	0.87	0.87	0.88	**0.91**	0.89
VLC2P	S@10	0.60	0.75	0.87*	0.86*	0.90*	0.88*	0.86	0.88	0.89	0.88	0.91	**0.92**
VLC2R	S@1	0.16	0.25	0.52*	0.48*	0.45*	**0.73***	0.48	0.55	0.62	0.59	0.59	0.68*
VLC2R	S@5	0.36	0.48	0.72*	0.69*	0.67*	**0.87***	0.67	0.71	0.75	0.75	0.73	0.74*
VLC2R	S@10	0.44	0.58	0.73*	0.72*	0.72*	**0.88***	0.72	0.73	0.75	0.75	0.74	0.76*

answer. (If that information were known in practice, perfection could easily be achieved by swapping the document at that position with the document at rank 1.)

4.1 Results

S@n results were computed for $n = 1, 5, 10$, for the baselines and for each of the six different reranking schemes on each of the five test sets. Full reranking and significance test results are tabulated in Tables II and III and highlighted graphically in Figures 10 and 11 in Appendix 1. We observed the following:

(1) All methods offered substantial improvements over the content baseline.
(2) All content rerankings significantly outperformed the random control.

Table III. Significant Differences Between Methods When Using Optimal Cutoffs (Each (nonrandom) method was compared against each of the others in turn and differences were tested for significance using the Wilcoxon test. Each significant difference found is shown with the direction of the difference.)

Collection	Type	Content	Anchor text
ANU	Popular	APR > DPR, URL	—
WT10gC	Random	DPR > Indeg APR > Indeg URL > Indeg, DPR, APR	URL > Indeg, DPR, APR
WT10gT	Random	Indeg > APR DPR > APR URL > Indeg, DPR, APR	APR > Indeg, DPR URL > Indeg, DPR, APR
VLC2P	Popular	—	APR > Indeg, DPR
VLC2R	Random	Indeg > APR URL > Indeg, DPR, APR	DPR > Indeg URL > Indeg, DPR, APR

(3) The only reranking method which showed significant benefit over the anchor text baseline was URL. This benefit was shown only for the *random* query sets. The benefits of reranking by URL were greatly diminished for anchor text compared to content baselines.
(4) URL performed at a consistently high level for both the content and anchor text baselines. The URL anchor text reranking was only outperformed in two cases: by APR on both ANU and VLC2P, cases where the query set and bookmarks were both derived from the same list of authoritative sources.
(5) For the *popular* home page queries (ANU and VLC2P), all anchor text rerankings outperformed their content counterparts.
(6) For *random* home page queries (WT10gT, WT10gC, and VLC2R), the content rerankings performed as well as, or better than, their anchor text counterparts.
(7) Improvements due to APR were only observed when using high-quality bookmarks, that is, when the query answers were to be found among the bookmarks.
(8) Improvements due to Indeg and DPR were almost identical.

5. REALISTIC COMBINATIONS

Many different schemes have been proposed for combining query-independent and query-dependent evidence, in the absence of unrealistic preknowledge. Kraaij et al. [2002] suggested measuring the query-independent evidence as a probability and treating it as a prior; however we used Okapi BM25 scores which are weights rather than probabilities. Westerveld et al. [2001] also made use of linear combinations of normalized scores, but for this to be useful with PageRank, a nonlinear transformation of the scores would almost certainly be needed. This is because while most PageRanks are very low a few are orders of magnitude larger. As shown in Figure 6, PageRank scores are distributed according to a power law.

Savoy and Rasolofo [2001] combined query-dependent URL length or URL similarity evidence with Okapi BM25 scores by reranking the top n documents on the basis of the URL scores. They also used data fusion techniques to improve on the results of individual combinations.

The Savoy and Rasolofo [2001] approach of reranking above a cutoff is consistent with our Optimal combinations and has been adopted for our Realistic experiments. In all experiments, we preserved the baseline ordering if the reranking scores were equal. We noted that such equality occurred more often in URL-type scores, which can take only one of four distinct values. To confirm that the superiority of URL-type reranking was not an artifact of quantization, we quantized[6] the hyperlink scores into four groups and observed a decrease in effectiveness. Hence we believe it is unlikely that URL-type has an unfair advantage due to this effect.

5.1 Setting Cutoffs

For the Realistic combinations, as opposed to the Optimal ones, we applied the same cutoff to all queries.

We considered two different strategies for choosing a suitable cutoff in the original ranking. In the first strategy (quota-based), the cutoff was set at $x\%$ of the number of documents retrieved for a query (maximum 1000). In the second (score-based), it was set at $x\%$ of the highest score for that query. In preliminary trials, we found that score-based cutoffs were far more effective than quota-based ones for all collections. Consequently, all reported results use score-based cutoffs.

We determined suitable score cutoffs for WT10gC by plotting S@5 against cutoff (see Figure 3) and recording the optimal cutoff for each reranking method. We then reranked using this cutoff on all other collections. Optimal cutoffs were calculated at S@5 due to the instability of P@1 and the smaller performance gains observed at S@10.

5.2 Results

Table IV and Figures 12 and 13 in Appendix 1 show the results of reranking the content and anchor text baselines using realistic cutoffs. From them, we observed that the following:

(1) URL reranking provided significant improvements over the anchor text and content baseline for WT10gT, VLC2P, and VLC2R. See Wilcoxon significance test results in Table IV.
(2) URL reranking performance was only surpassed by APR on the ANU collection (at S@1) where APR used very high-quality bookmarks.
(3) None of the hyperlink based schemes provided a significant improvement over the anchor text baseline.
(4) For the *popular* query sets (ANU and VLC2P), the anchor text baseline with URL reranking produced the best performance.
(5) For the *random* query sets (WT10gT and VLC2R), the content baseline with URL reranking produced the best performance.

[6]Grouped similar scores to reduce the number of possible values.

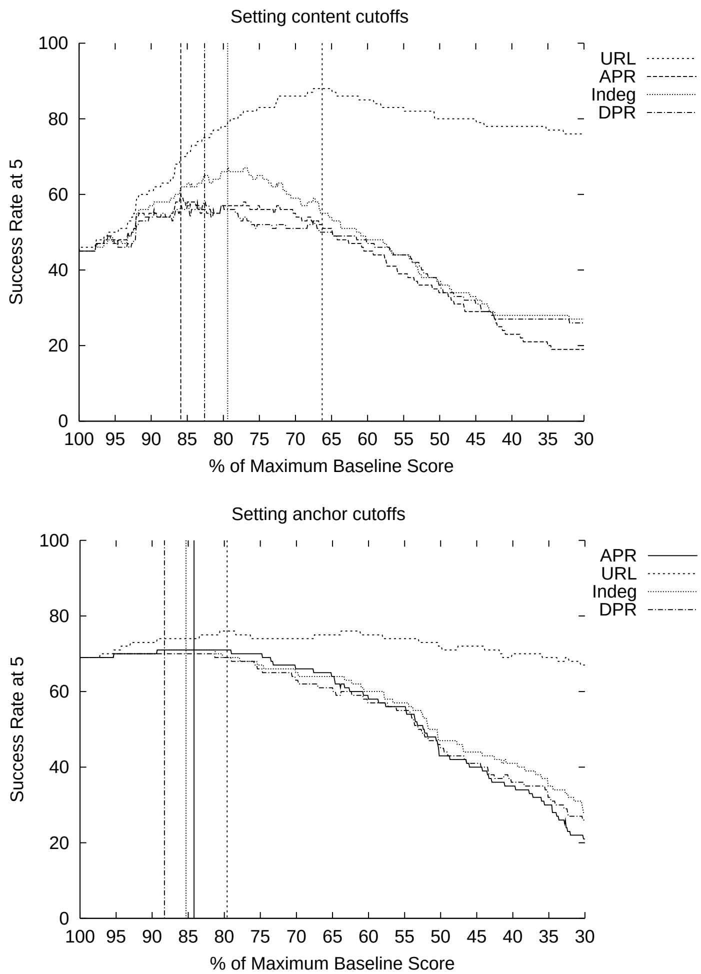

Fig. 3. Setting score cutoffs for rerankings of the content (top) and anchor text (bottom) baselines using the WT10gC collection. The vertical lines represent the chosen cutoff values, which were then used in all realistic experiments. Note that if the optimal cutoff spanned multiple values, we used the mean of those values. Numerical cutoff scores are provided in Table IV.

(6) APR decreased retrieval performance for both baselines on the VLC2R collection (APR on VLC2R was influenced by Yahoo! but queries were not).

6. DISCUSSION

In this section we discuss how to choose the best combination of baseline and query-independent evidence.

6.1 What Query-Independent Evidence Should Be Used?

The Optimal combination results show that rerankings by all of the query-independent methods considered were significantly better than the random control for the content baseline. Further, for all *random* query sets, URL reranking

Table IV. Realistic Cutoff Results (The Realistic combination experiment is described in Section 5. The best combinations are highlighted in bold. Cutoffs (shown as "(at X)") were obtained by training on WT10gC at S@5 (represented in italics). "Sig." reports the statistical significance of the improvements ($>$ 0.05 is significant). Significance was tested using the Wilcoxon matched-pairs signed ranks test. An "N" indicates the observed improvements were not significant.)

		Content					Anchor text				
Coll.	Meas.	Base	Indeg (at 20.6)	DPR (at 17.4)	APR (at 14.1)	URL (at 33.7)	Base	Indeg (at 14.7)	DPR (at 11.7)	APR (at 15.8)	URL (at 20.4)
ANU	S@1	0.28	0.36	0.27	0.48	0.39	0.76	0.75	0.77	**0.81**	0.78
ANU	S@5	0.50	0.60	0.56	0.67	0.73	0.96	0.95	0.94	0.95	**0.98**
ANU	S@10	0.58	0.73	0.69	0.72	0.83	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
ANU	Sig.	—	N	N	**0.00**	**0.00**	—	N	N	N	N
WT10gC	S@1	0.23	0.36	0.38	0.33	**0.71**	0.48	0.52	0.51	0.52	0.67
WT10gC	S@5	*0.45*	*0.67*	*0.58*	*0.60*	***0.88***	*0.69*	*0.71*	*0.70*	*0.71*	*0.76*
WT10gC	S@10	0.55	0.73	0.67	0.66	**0.90**	0.72	0.72	0.72	0.72	0.76
WT10gT	S@1	0.22	0.46	0.41	0.32	**0.70**	0.54	0.52	0.53	0.48	0.65
WT10gT	S@5	0.48	0.64	0.59	0.62	**0.83**	0.68	0.70	0.69	0.70	0.73
WT10gT	S@10	0.59	0.71	0.69	0.66	**0.88**	0.72	0.73	0.72	0.73	0.74
WT10gT	Sig.	—	N	N	N	**0.00**	—	N	N	N	**0.00**
VLC2P	S@1	0.27	0.38	0.37	0.39	0.56	0.69	0.68	0.69	0.72	**0.79**
VLC2P	S@5	0.51	0.60	0.56	0.61	0.68	0.85	0.83	0.83	0.84	**0.88**
VLC2P	S@10	0.60	0.69	0.66	0.75	0.76	0.86	0.85	0.87	0.85	**0.90**
VLC2P	Sig.	—	N	N	**0.01**	**0.01**	—	N	N	N	**0.00**
VLC2R	S@1	0.16	0.25	0.20	0.21	0.63	0.48	0.47	0.46	0.41	**0.66**
VLC2R	S@5	0.36	0.48	0.44	0.44	**0.82**	0.67	0.70	0.70	0.68	0.73
VLC2R	S@10	0.44	0.57	0.52	0.53	**0.83**	0.72	0.73	0.72	0.69	0.76
VLC2R	Sig.	—	N	N	N	**0.00**	—	N	N	N	**0.00**

Table V. Numerical Summary of Improvements ("Sig." denotes whether the improvements were shown to be significant using the Wilcoxon test. The percentile realistic improvements were calculated as a percentage improvement over the best baseline (which was anchor text in every case). "AT+*" denotes a combination of anchor text with any of the query-independent evidence examined here. "AT+URL" denotes a combination of anchor text with URL-type query-independent evidence. "AT+APR" denotes a combination of anchor text with APR query-independent evidence. "C+URL" denotes a combination of content with URL-type query-independent evidence.)

Collection Info			Optimal	Realistic				
Coll.	Type	B'mark Quality	Best S@5	Best S@5	S@1 Improve	S@5 Improve	S@10 Improve	Sig.
ANU	Pop.	v.High	0.98 AT+*	0.98 AT+URL	*2.6%* 0.76→0.78	*2.0%* 0.96→0.98	*0%* 0.98→0.98	No
WT10gT	Rand.	Low	0.88 C+URL	0.83 C+URL	*23.9%* 0.54→0.70	*18.1%* 0.68→0.83	*18.2%* 0.72→0.88	Yes
VLC2P	Pop.	High	0.91 AT+APR	0.88 AT+URL	*13.7%* 0.69→0.79	*3.4%* 0.85→0.88	*4.3%* 0.86→0.90	Yes
VLC2R	Rand.	Low	0.87 C+URL	0.82 C+URL	*14.8%* 0.48→0.63	*18.3%* 0.67→0.82	*13.3%* 0.72→0.83	Yes

of the anchor text baseline was significantly better than the random control. Results were quite stable across collections despite differences in their scale.

Naturally, the benefits of reranking above realistic cutoffs were smaller, but the URL method in particular achieved substantial gains over both baselines, as reported in Table V. It is clear that classification of URL-types is of considerable value in a home page finding system.

It is of interest that URL reranking results for the ANU collection were rather poorer than for the other collections. Although investigation confirmed UTwente/TNO's ordering, that is root(36/137) > subroot (50/862) > directory (72/14,059) > file (40/382,274),[7] the ratio for the URL subroot class was higher than for other collections.

It should be noted that URL reranking would be of little use in Webs in which URLs exhibit no hierarchical structure. Some organizations publish URLs of the form `xyz.org/getdoc.cgi?docid=9999999` and in this case there are no subroots or paths.

Hyperlink recommendation results indicate these schemes may have relatively little role to play in home page finding tasks for collections within the range of sizes studied here (400,000 to 18.5 million pages). While Optimal reranking improvements over the content baseline were encouraging, the performance improvements over the anchor text baseline were minimal. This suggests that most of the potential improvement offered by hyperlink recommendation methods is already exploited by the anchor text baseline. In most of the Realistic rerankings, it was almost impossible to differentiate between the reranking of the anchor text baseline and the baseline itself. Throughout our experiments in-degree appeared to provide the most consistent performance improvement. APR performed well when using high-quality bookmark sets but degraded performance when using poor bookmark sets on *random* (WT10gT and VLC2R) query sets. The improvement achieved by these methods relative to the anchor text baselines was not significant (see Table IV).

The results for the two different versions of PageRank show that PageRank's contribution to home page finding on collections of this size is very dependent upon the choice of bookmark pages. However, even for *popular* queries (ANU and VLC2P) APR results are generally inferior to those for URL rerankings. Further, the results for APR are not much better than for DPR. Of the three hyperlink recommendation methods, in-degree may be the best choice, as the PageRank variants offer little advantage and are more computationally expensive.

6.2 Which Baseline Should Be Used?

Before reranking, the anchor text baseline always outperforms the content, by 28–45% (see Figure 3). However, on two of the four collections, URL-type rerankings of content outperform similar rerankings of anchor text. These two cases are the ones for which the target home pages were *randomly* chosen. The effect was not observed for the *popular* targets.

Figure 4 illustrates the difference between the *random* and *popular* sets by plotting S@N against N for both baselines. For the *popular* query set, the two baselines converge at about $N = 500$, but for the *random* set the content baseline is clearly superior for $N > 150$. The plot for VLC2R is similar to that observed in a previous study of content and anchor text performance on the WT10gT collection [Kraaij et al. 2002]. The explanation of the effect is believed to be as follows.

[7]Note that in these figures all URLs (including equivalent URLs) were considered.

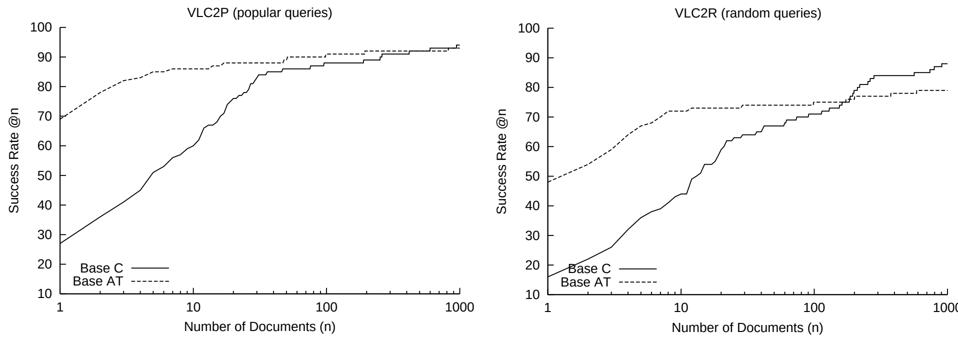

Fig. 4. Success rates across different cutoffs. The left plot is for VLC2P, the VLC2 crawl with a *popular* home page query set. The right plot is for VLC2R, the same crawl but with a *random* home page query set. The anchor text baseline performs well between 0–150 for both collections. In VLC2P, at around S@150 the anchor text baseline performance approaches the content baseline performance. In VLC2R the anchor text performance is surpassed by the content performance at around S@150. These plots are consistent with the S@1000 values reported in Table I.

Even though anchor text rankings are better able to discriminate between home pages and other relevant pages, full anchor text rankings are shorter[8] than those for content. Some home pages have no useful incoming anchor text and therefore do not appear anywhere in the anchor text ranking. By contrast, most home pages do contain some form of site name within their content and will eventually appear in the content ranking.

Selecting queries from a directory within the collection guarantees that the anchor document for the target home page will not be empty, but there is no such guarantee for randomly chosen home pages. Selection of home pages for listing in a directory is undoubtedly biased toward useful, important, or well-known sites which are also more likely to be linked to from other pages. It should be noted that incoming home page queries would probably also be biased toward this type of site.

7. FURTHER EXPERIMENTS

Having established the principal results above, we conducted a series of followup experiments. In particular we investigated the following:

— to what extent results can be understood in terms of rank and score distributions;
— to what extent PageRank effectiveness is dependent on tuning the d value;
— whether other classifications of URL-type provide similar, or superior, performance;
— to what extent our PageRanks and in-degrees correlated with those reported by Google; and
— whether the use of anchor text and link graph information external to the collection improved retrieval effectiveness

[8]Ignoring documents achieving a zero score.

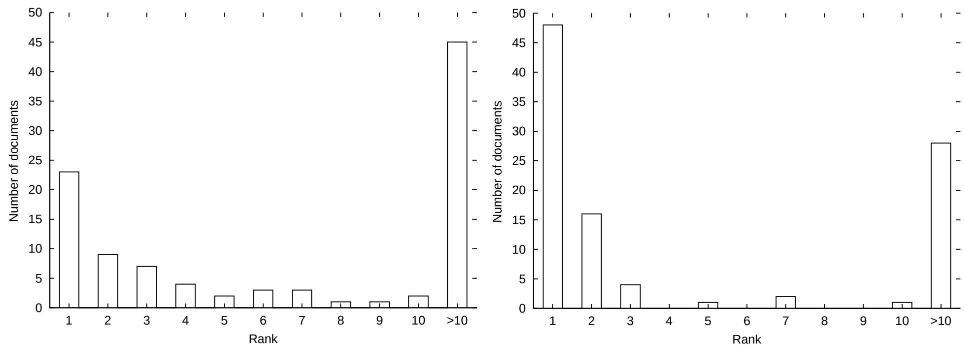

Fig. 5. Baseline rankings of the correct answers for WT10gC (content left, anchor text right). The correct answer was retrieved within the top 10 results for over 50% of queries on both baselines. The anchor text baseline has the correct answer ranked as the top result on almost 50% of the queries.

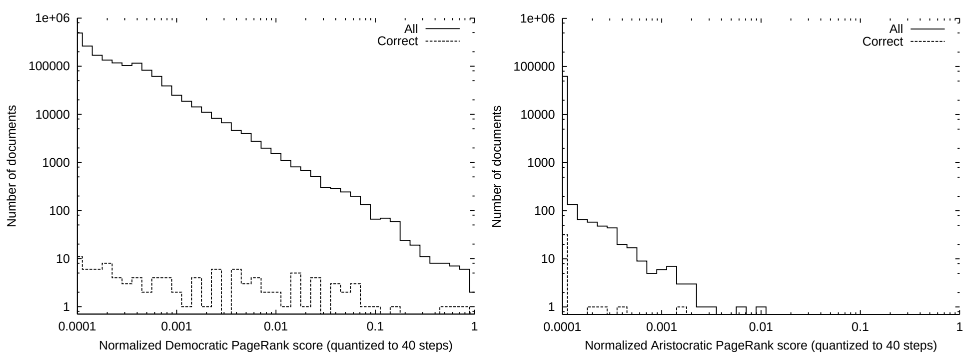

Fig. 6. PageRank distributions for the WT10gC collection (DPR left, APR right). These plots contain the distribution of all pages in the collection (All) and the distribution of the 100 correct answers (Correct). The distribution of the DPR scores for all pages follow the power law. In contrast, the correct answers are spread more evenly across Democratic PageRank scores. The proportion of pages which are correct answers increases at higher PageRanks. There are many pages which do not achieve an APR score; thus merely having an APR score > 0 is a good indicator of a page being a correct answers.

7.1 Rank and Score Distributions

Here we present an analysis of the distribution of correct answers for each type of evidence over the WT10gC collection.

The baseline rankings of the correct answers are plotted in Figure 5. In over 50% of occasions both the content and anchor text baselines contain the correct answer within the top 10 results. Anchor text provides the better scoring of the two baselines, with the correct home page ranked as the top result for almost 50% of the queries. This demonstrates the effectiveness of anchor text as a home page finding measure (as shown previously in Craswell et al. [2001] and Bailey et al. [2003]).

The PageRank distributions are plotted in Figure 6. The distribution of the Democratic PageRank scores for all pages follows the power law. In contrast,

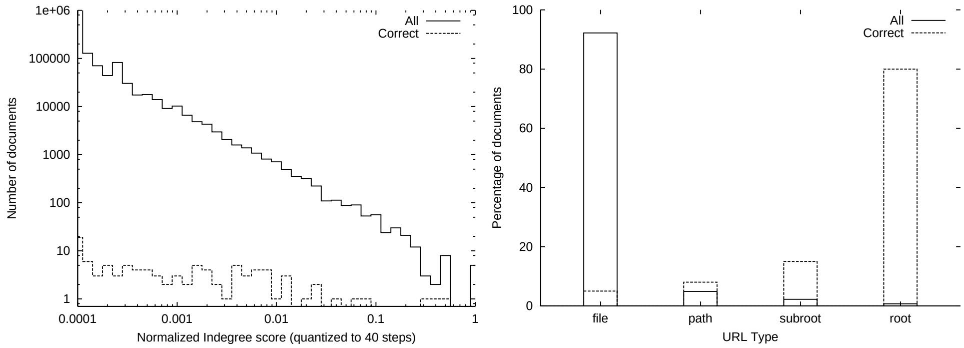

Fig. 7. Other distributions for the WT10gC collection (in-degree left, URL-type right). The left plot contains the in-degree distribution for all pages (All) and the 100 correct answers (Correct). The distribution of the in-degree scores for all pages follow the power law. In contrast, the correct answers are spread more evenly across in-degree scores. The proportion of pages which are correct answers increases at higher in-degree scores. The right plot contains the URL-type distribution (in percentages) of all pages (All) and the correct answers (Correct). The "root" tier contains only 1% of the pages in the collection, but 80% of the correct answers. In contrast, the "file" tier contains 92% of the collections pages but only 5% of the correct answers.

the distribution of correct answers is spread, with the proportion of pages that are correct answers increasing at higher PageRanks. There are many pages which do not achieve an APR score; thus merely having an APR score > 0 is a good indicator of a page being a correct answer. These plots indicate that both forms of PageRank provide some sort of home page evidence.

The in-degree distribution is plotted on the left in Figure 7 and is similar to the Democratic PageRank distribution. In comparison, the graph is slightly skewed to the left, indicating that there are more pages with low in-degrees than there are pages with low PageRanks. The distribution of correct answers is spread across in-degree scores, with the proportion of pages that are correct answers increasing at higher in-degrees. This shows in-degree also provides some sort of home page evidence.

The URL-type distribution is plotted on the right in Figure 7. URL-type is a useful home page indicator with a large proportion of the correct answers located in the "root" tier and few correct answers located within the "file" tier.

7.2 How Sensitive is PageRank to *d* Value Modifications?

Figure 8 shows how the performance of PageRank on the WT10gC collection is affected by changes to the d value. Figure 9 reports the number of PageRank iterations performed at each corresponding d value.

We observed that the performance of PageRank was remarkably stable even with large changes to the d value. When we set $d = 0.02$, the performance of the optimal reranking was similar to the performance at $d = 0.85$. Without the introduction of any random noise (at $d = 1.0$), the PageRank calculation did not converge. The PageRank calculation did converge when we introduced only a small amount of random noise (setting $d = 0.99$).

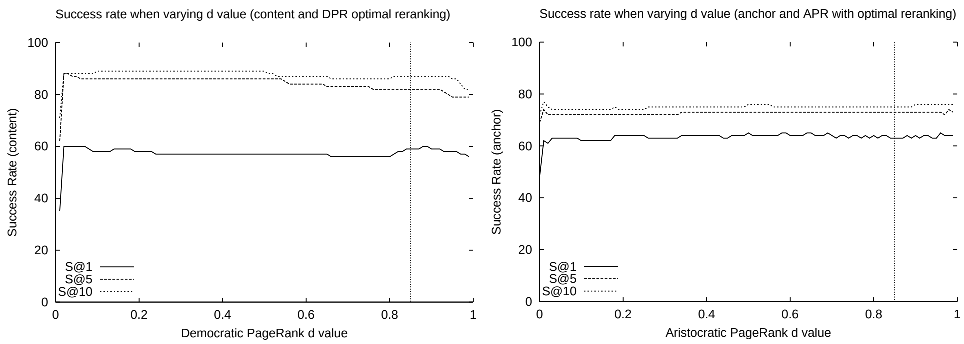

Fig. 8. d value variations for PageRank calculations over the WT10g collection (DPR left, APR right). As d approaches 0, the bookmarks become more influential. As d approaches 1, the calculation approaches "pure" PageRank (i.e., a PageRank calculation with no random jumps).

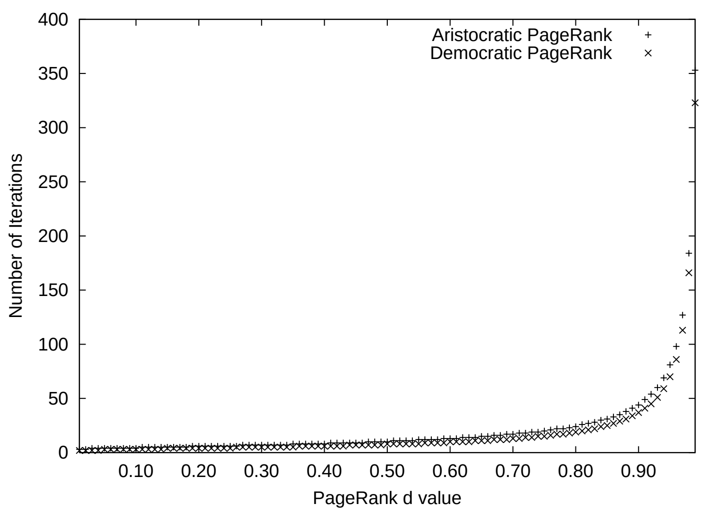

Fig. 9. Number of PageRank Iterations when varying the d value over WT10g. PageRank did not converge at $d = 1$ (no random jumps).

As we observed little improvement in performance when increasing the d value, to minimize the computation required for PageRank, d should be set to around 0.05 for collections of this size.

7.3 Are the Combinations in the Four-Tier URL-Type Classification Optimal?

Here we evaluate how combining the four URL-type classes and introducing length- and directory depth-based scores changes retrieval effectiveness. The most effective scoring methods evaluated are presented in Table VI.

None of the new URL-type methods significantly improved upon the performance of the original URL-type classes (root > subroot > directory > file). However, we found that combining the "subroot" and "directory" classes did not adversely affect URL-type effectiveness. We also obtained good performance

Table VI. S@5 for URL-Type Category Combinations (Length (how long a URL is, favoring short directories) and directory depth (how many directories the URL contains, favoring URLs with shallow directories). R represents the "root" tier, S represents the "subroot" tier, D is for the "directory" tier, and F is for the "file" tier.)

Dataset	Baseline	Original (R>S>D>F)	Length	Directory Depth	root>other (R>S+D+F)	root>dir>file (R>S+D>F)	Dir len. >file (R>S>D(len.)>F)
ANU	Content	87	88	68	62	77	88
ANU	Anchor text	98	98	98	96	98	98
WT10gC	Content	89	91	72	83	89	91
WT10gC	Anchor text	81	83	74	77	81	81
WT10gT	Content	87	88	73	80	85	87
WT10gT	Anchor text	82	83	80	80	82	82
VLC2P	Content	87	85	68	81	85	87
VLC2P	Anchor text	89	91	87	88	88	89
VLC2R	Content	87	86	62	82	85	87
VLC2R	Anchor text	74	76	73	73	74	74

Table VII. Correlation of PageRank Variants with In-degree (Pearson r.); All Were Significant at the 0.05 Level

	DPR	APR	No. of pages (millions)
ANU	0.836	0.448	0.40
WT10g	0.71	0.555	1.69
VLC2	0.666	0.164	18.57

using a simple URL length measure. Here pages were ranked according to the length in characters of their URL (favoring short URLs). "File" URLs contain filenames and are thereby longer than their "root" and "directory" counterparts. This may explain the good performance of the URL length measure. Reranking baselines using only the URL directory depth (number of slashes in the URL) performed relatively poorly.

We conclude that when using URL-type scores for home page finding tasks it is important to distinguish between "root," "directory," and "file" pages. This can be done either explicitly through a categorization of URL-types or by measuring the length of the URL.

7.4 PageRank Correlations

Table VII shows that DPR and in-degree are highly correlated but that the correlation tends to weaken as the size of the collection increases. This weaker association as collection size increases suggests that PageRank might have quite different properties for very large crawls. Google's PageRank, based on 50–100 times more documents than are in VLC2, is likely to be different and possibly superior to the PageRanks studied here. In addition, Google may use a different PageRank variant and different bookmarks.

To further understand the PageRank employed by the main Google Web search engine, we compared our PageRank scores with the Google PageRanks

Table VIII. Hybrid WT10g/VLC2 Run Results (Note that the WT10g collection is a subset of the VLC2 collection. The WT10g anchor text scores are the baselines used throughout all other experiments. The VLC2 anchor scores are new rankings that use external anchor text from the VLC2 collection. WT10g DPR is a Democratic PageRank reranking using the link table from the WT10g collection. VLC2 DPR is a Democratic PageRank reranking using the link table from the VLC2 collection. The use of the (larger) VLC2 link table DPR scores did not significantly improve the performance of DPR reranking. The use of external anchor text, taken from the VLC2 collection, provided significant performance gains.)

	WT10g anchor text			VLC2 anchor text		
	—	DPR WT10g	DPR VLC2	—	DPR WT10g	DPR VLC2
WT10gC	0.69	0.72	0.69	0.78	**0.79**	0.78
WT10gT	0.68	0.72	0.72	0.72	0.72	**0.73**

reported for all 201 ANU pages listed in the Google Directory.[9] For those pages, PageRanks were extracted from Google's DMOZ directory and in-degrees were extracted using the Google `link:` query operator. Google PageRank and in-degree were correlated ($r = 0.358$), as they were for ANU, WT10g, and VLC2. Also, the correlation between Google in-degree and ANU in-degree was very strong ($r = 0.933$). Google's in-degrees, based on a much larger crawl, were only three times larger than those from the ANU crawl.

While Google PageRank and ANU PageRank correlated over the 201 observations, the correlation was less strong than for in-degree (DPR $r = .26$, APR $r = .31$). This again indicates that Google PageRank is different from the PageRanks studied here. Note that only five different values of PageRank were reported by Google for the 201 pages (11, 16, 22, 27, and 32 out of 40). Google PageRanks made available to the public might be different (transformed and quantized) from those used in its internal ranking.

Although this study is not directly applicable to very large crawls, its results are quite stable for a range of smaller multiserver crawls. The range of sizes of our collections (400,000 to 18.5 million pages) is typical of many enterprise webs and thus is interesting both scientifically and commercially.[10]

7.5 Use of External Link Information

To explore the effects of increasing collection size, we performed a series of hybrid WT10g/VLC2 runs. This is potentially revealing because the WT10g collection is a subset of the VLC2 collection. The runs, shown in Table VIII, used combinations of WT10g corpus data and VLC2 link information. Our hypotheses were that by using link tables from the larger collection we would obtain a more complete link graph and thereby improve the performance of the hyperlink recommendation methods. Further, we expected an improvement in anchor text performance (due to a potential increase in the amount of available

[9]A version of the manually constructed DMOZ Open Web directory [DMOZ n.d.] which reports Google PageRanks. The Google DMOZ Directory is available at `http://directory.google.com` [Google 2000].

[10]The rated capacities of the two Google search appliances are in fact very similar to these sizes (150,000 and 15 million pages).

anchor text). Note that during these hybrid runs we removed all VLC2 anchor text that pointed to pages outside the WT10g collection.

Surprisingly, the use of the (larger) VLC2 link table DPR scores did not noticeably improve the performance of DPR reranking. However, the use of external anchor text, taken from the VLC2 collection, provided significant performance gains. This would suggest that in situations where an enterprise or small web has link information for a larger web, benefits will be seen if the anchor text from the external link graph is recorded and used for the smaller collection.

We note that WT10g is not a uniform sample of VLC2 but was engineered to maximize the interconnectivity of the documents selected [Bailey et al. 2003]. Hence, the effects of scaling up may be smaller than would be expected.

8. CONCLUSIONS

Reranking query-dependent baselines (both content and anchor text) on the basis of URL-type produced consistent benefits. This heuristic would be a valuable component of a home page finding system for web collections with explicit hierarchical structure.

By contrast, outside our optimal experiments we have yet to achieve any significant performance improvement through the use of hyperlink-based recommendation schemes. Even on the WT10gC collection, on which the reranking cutoffs were trained, the recommendation results were disappointing. For collections of less than 20 million pages, the hyperlink recommendation methods do not seem to provide practical benefits on a home page finding task. Similarly, few benefits has been found on relevance-based retrieval in the TREC Web Track [Hawking et al. 1999]. Further work is required to determine whether such schemes could be useful in other tasks such as topic distillation.

An ideal home page finding system would be able to exploit both anchor text (for superior performance when targeting popular sites) and content information (to ensure that home pages with inadequate anchor text are not missed). Further work is needed to determine how to optimally combine all sources of evidence for the home page finding task and how to provide best all round search effectiveness when home page queries are interspersed with other query types.

APPENDIX

A.1 Graphical Representations of Rerankings

(See Figures 10, 11, 12 and 13.)

A.2 Okapi BM25 Formula

The content and anchor text baselines in this paper use the BM25 formula derived by Robertson et al. [1994] with the parameters ($k_1 = 2.0, k_2 = 0.0, k_3 = \infty, b = 0.75$).

$$w_t = q_t \times tf_d \times \frac{\log\left(\dfrac{N - n + 0.5}{n + 0.5}\right)}{2 \times \left(0.25 + 0.75 \times \dfrac{dl}{avdl}\right) + tf_d} \tag{1}$$

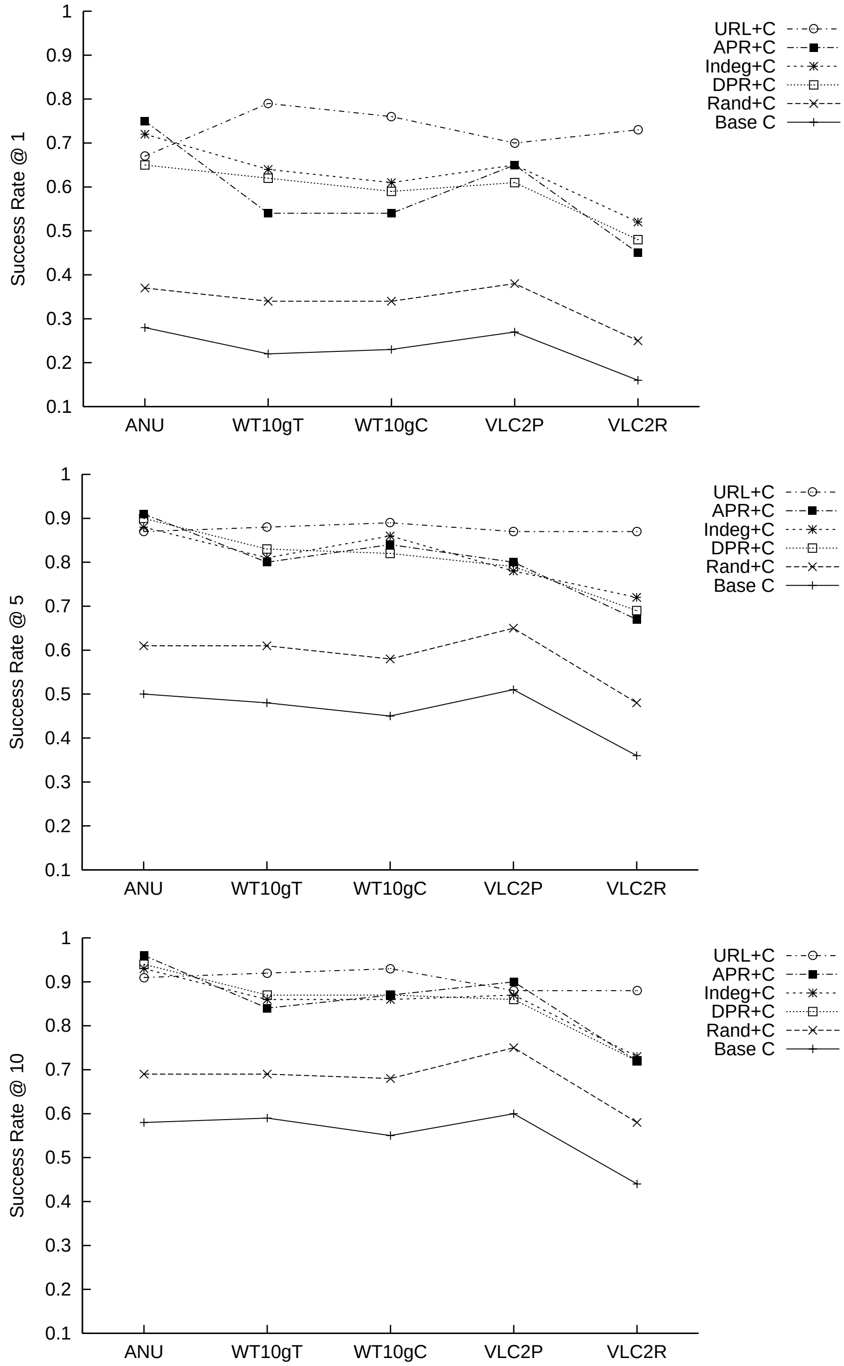

Fig. 10. S@1, S@5, and S@10 results for reranking above an optimal cutoff against the content baseline.

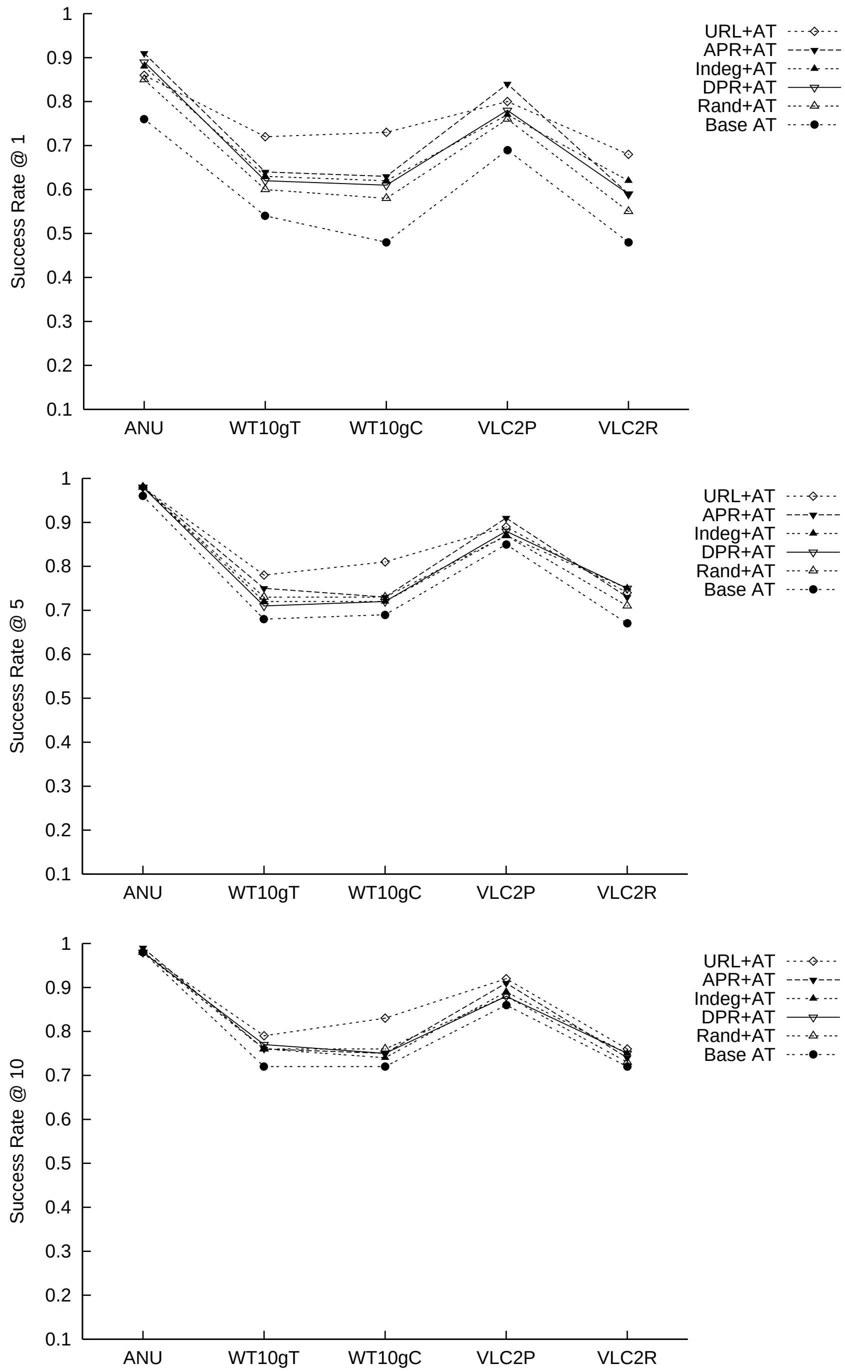

Fig. 11. S@1, S@5, and S@10 results for reranking above an optimal cutoff against the anchor text baseline.

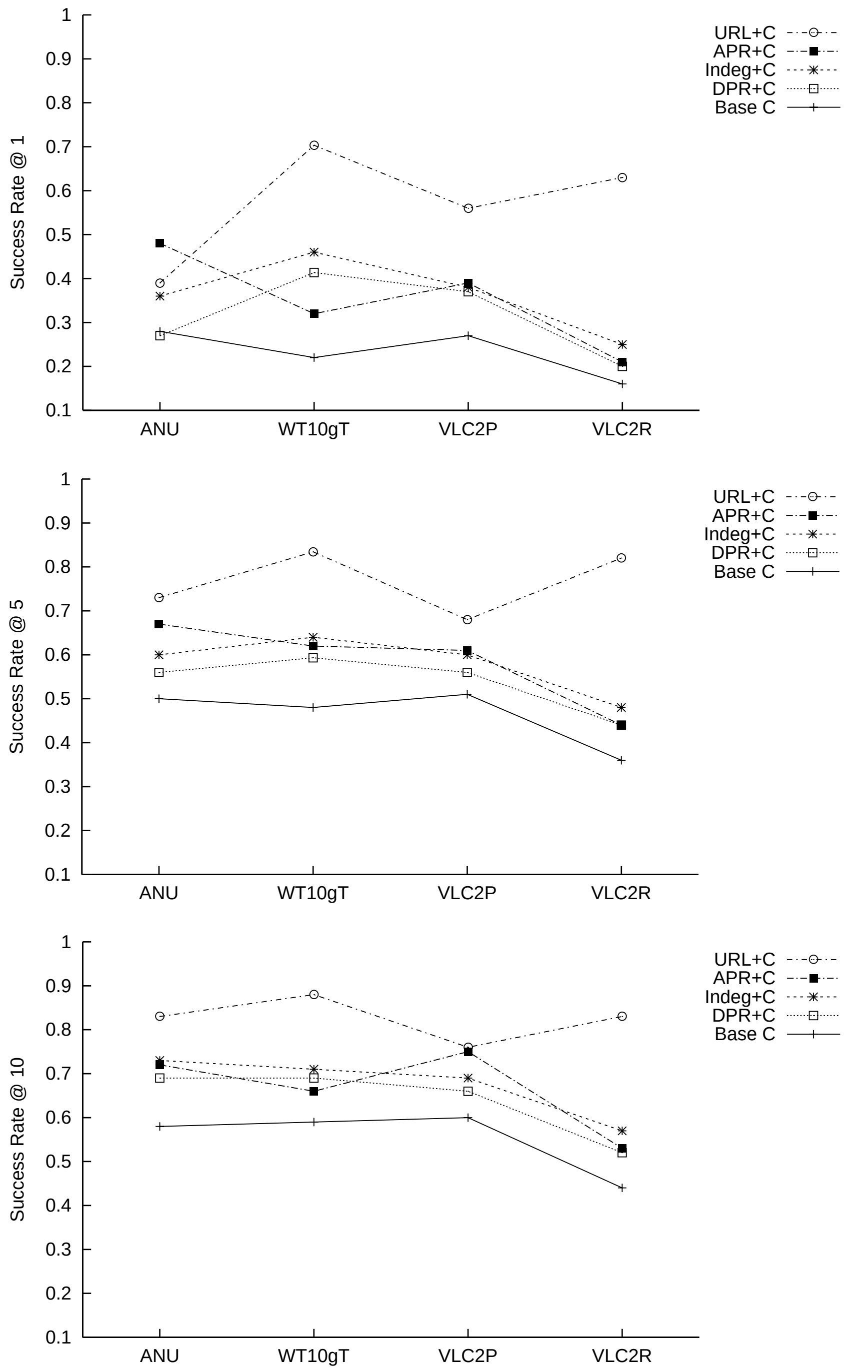

Fig. 12. S@1, S@5, and S@10 results for reranking above a realistic cutoff against the content baseline. Note that WT10gC is not reported in these graphs as it was used as training data. The results for WT10gC are presented in Table IV.

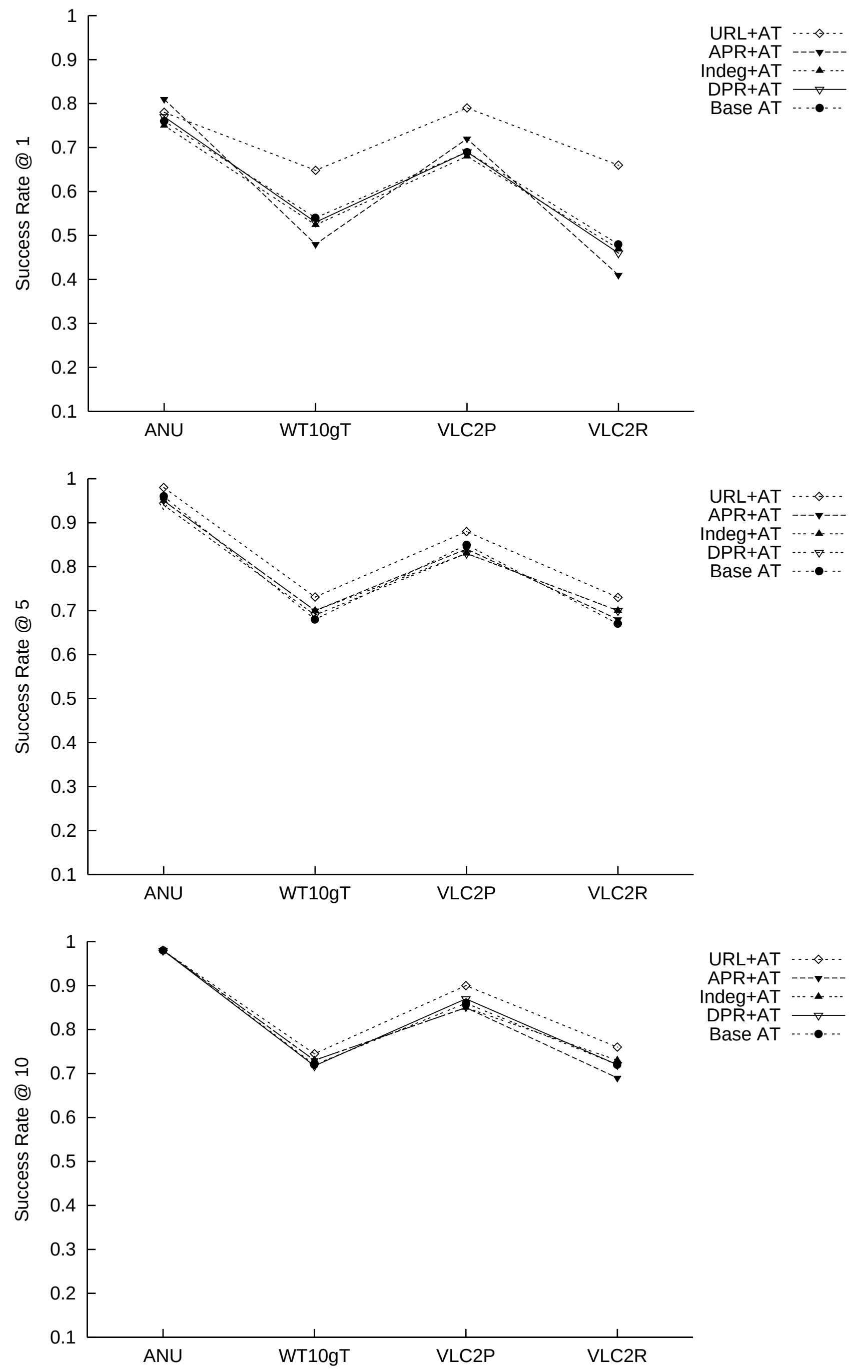

Fig. 13. S@1, S@5, and S@10 results for reranking above a realistic cutoff against the anchor text baseline. Note that WT10gC is not reported in these graphs as it was used as training data. The results for WT10gC are presented in Table IV.

where w_t is the relevance weight assigned to a document due to query term t, q_t is the weight attached to the term by the query, tf_d is the number of times t occurs in the document, N is the total number of documents, n is the number of documents containing at least one occurrence of t, dl is the length of the document, and $avdl$ is the average document length (both measured in bytes).

REFERENCES

AMENTO, B., TERVEEN, L. G., AND HILL, W. C. 2000. Does "authority" mean quality? Predicting expert quality ratings of Web documents. In *Proceedings of ACM SIGIR'00* (Athens, Greece). 296–303.

BAILEY, P., CRASWELL, N., AND HAWKING, D. 2003. Engineering a multi-purpose test collection for Web retrieval experiments. *Inform. Process. Manage.* To appear. Available online at `http://pigfish.vic.cmis.csiro.au/~nickc/pubs/cwc.ps.gz`.

BHARAT, K. AND MIHAILA, G. A. 2001. When experts agree: Using non-affiliated experts to rank popular topics. In *Proceedings of WWW10* (Hong Kong). Available online at `www10.org/cdrom/papers/474/`.

BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of WWW7* (Brisbane). 107–117. Available online at `www7.scu.edu.au/programme/fullpapers/1921/com1921.htm`.

BRODER, A. 2002. A taxonomy of web search. *SIGIR For. 36,* 2 (Fall), 3–10.

CARRIÈRE, S. J. AND KAZMAN, R. 1997. WebQuery-Searching and visualizing the Web through connectivity. In *Proceedings of WWW6* (Santa Clara, CA). 701–711. Available online at `www.scope.gmd.de/info/www6/technical/paper096/paper96.html`.

CRASWELL, N., HAWKING, D., AND ROBERTSON, S. 2001. Effective site finding using link anchor information. In *Proceedings of ACM SIGIR'01* (New Orleans, LA). 250–257.

DMOZ. n.d. Open directory project. See Web site `www.dmoz.org`.

EXCITE. 2002. Excite. See Web site `www.excite.com`.

FAST SEARCH AND TRANSFER, ASA. Personal communication. See Web site `www.alltheweb.com`.

GOOGLE. 2003a. Google search appliance frequently asked questions. Available online at `www.google.com/appliance/faq.html`.

GOOGLE. 2003b. Google search engine. See Web site `www.google.com`.

HAWKING, D. 2000. Overview of the TREC-9 Web Track. In *Proceedings of TREC-9*. Available online at `trec.nist.gov/pubs/trec9/`.

HAWKING, D. AND CRASWELL, N. 2001. Overview of the TREC-2001 Web Track. In *Proceedings of TREC-2001* (Gaithersburg, MD). Available online at `trec.nist.gov/pubs/`.

HAWKING, D., CRASWELL, N., AND GRIFFITHS, K. 2001. Which search engine is best at finding online services? In *WWW10 Poster Proceedings* (Hong Kong). Available online at `www10.org/cdrom/posters/1089.pdf`.

HAWKING, D., VOORHEES, E., BAILEY, P., AND CRASWELL, N. 1999. Overview of TREC-8 Web Track. In *Proceedings of TREC-8*. 131–150. Available online at `trec.nist.gov/pubs/trec8/`.

KRAAIJ, W., WESTERVELD, T., AND HIEMSTRA, D. 2002. The importance of prior probabilities for entry page search. *Proceedings of ACM SIGIR 2002*.

Looksmart. n.d. See Web site `www.looksmart.com`.

NG, A. Y., ZHENG, A. X., AND JORDAN, M. I. 2001. Link analysis, eigenvectors, and stability. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence* (IJCAI-01). ACM Press, New York, NY.

PAGE, L., BRIN, S., MOTWANI, R., AND WINOGRAD, T. 1998. The PageRank citation ranking: Bringing order to the Web. Tech. Rep. 1999-66, Stanford University Database Group. Stanford, CA. Available online at `dbpubs.stanford.edu:8090/pub/1999-66`.

ROBERTSON, S., WALKER, S., JONES, S., HANCOCK-BEAULIEU, M., AND GATFORD, M. 1994. Okapi at TREC-3. In *Proceedings of TREC-3*. Available online at `trec.nist.gov/pubs/trec3/`.

SAVOY, J. AND RASOLOFO, Y. 2001. Report on the TREC-10 experiment: Distributed collections and entrypage searching. In *TREC-2001 Notebook proceedings*. Available online at `trec.nist.gov/pubs/`.

SINGHAL, A. AND KASZKIEL, M. 2001. A case study in Web search using TREC algorithms. In *Proceedings of WWW10* (Hong Kong). 708–716. Available online at www10.org/cdrom/papers/317/.

TRAVIS, B. AND BRODER, A. 2001. Web search quality vs. informational relevance. In *Proceedings of the Infonortics Search Engines Meeting*. Boston. Available online at www.infonortics.com/searchengines/sh01/slides-01/travis.html.

WESTERVELD, T., KRAIJ, W., AND HIEMSTRA, D. 2001. Retrieving Web pages using content, links, URLs and anchors. In *TREC-2001 Notebook Proceedings*. Available online at trec.nist.gov/pubs/.

YAHOO!. n.d. Yahoo! directory service. See Web site www.yahoo.com.

Received August 2002; revised February 2003; accepted May 2003