Effective Site Finding using Link Anchor Information

Nick Craswell and David Hawking
CSIRO Mathematical and Information Sciences
Canberra, Australia
{Nick.Craswell,David.Hawking}@cmis.csiro.au

Stephen Robertson
Microsoft Research
Cambridge, UK
ser@microsoft.com

ABSTRACT

Link-based ranking methods have been described in the literature and applied in commercial Web search engines. However, according to recent TREC experiments, they are no better than traditional content-based methods. We conduct a different type of experiment, in which the task is to find the main entry point of a specific Web site. In our experiments, ranking based on link anchor text is twice as effective as ranking based on document content, even though both methods used the same BM25 formula. We obtained these results using two sets of 100 queries on a 18.5 million document set and another set of 100 on a 0.4 million document set. This site finding effectiveness begins to explain why many search engines have adopted link methods. It also opens a rich new area for effectiveness improvement, where traditional methods fail.

Keywords

Citation and Link Analysis, Web IR, Evaluation

1. INTRODUCTION

In principle, if a Web site exists, it should be possible for a user to find it. However, manually maintaining a directory of all Web sites is difficult, because of the Web's size and volatility. For this reason, effective automatic site finding is an interesting research problem.

We compare the site finding effectiveness of a link-based ranking method and a content-based ranking method. Link methods are those which make some use of the hypertext structure of the Web. Using link methods, a document's ranking is based (at least in part) on its incoming and outgoing links. Content methods are more traditional, where a document's ranking is based on its text content.

Past TREC experiments have found that link information does not enhance retrieval effectiveness. In particular, TREC-8 Small and Large Web Tracks [7] found link methods to be no better than non-link methods. These negative results may seem surprising, given the enthusiasm of researchers [3, 8] and industry [12] for link methods.

Our new experiments are also based on TREC methodology, with a static test collection and blind examination of pooled results. However, TREC experiments have been based on a subject search task, while these experiments are based on a site finding task. Furthermore, we use a link anchor method [9, 3], propagating link anchor words from the source to the target document. TREC experiments have focused on other link methods (see Section 3). To our knowledge, these are the first TREC-style experiments to evaluate site finding effectiveness and the first to show a link-based effectiveness enhancement.

We describe the site finding problem (Section 2) and link-based ranking methods (Section 3). Next we introduce an evaluation methodology for evaluating site finding effectiveness (Section 4). Then we present our site finding experiments, comparing the effectiveness of a link method and a non-link method, followed by discussion and conclusions (Sections 5–7).

2. THE SITE FINDING PROBLEM

It is difficult to precisely and uncontroversially define the term "Web site". Rather than introduce a new definition, we adopt one by Amento et al [1]:

> A site (multimedia document) is an organized collection of pages on a specific topic maintained by a single person or group. . . . A site is not the same thing as a domain: for example, thousands of sites are hosted on www.geocities.com.

We note that the "topic" of a site might be quite broad. For example, we would consider Yahoo! (http://www.yahoo.com/) to be a site, but it covers a broad range of subject matter and provides a range of services. We also note that an important feature of a site is that its pages are published together, as a coherent source of information.

SIGIR 2001 has a Web site and many of its paper authors also publish their own sites. Included in this definition are sites which are a subset of a larger site. For example, the Microsoft Word site is part of the larger Microsoft site.

Most sites have a main entry page, sometimes also referred to as a home page. This page usually has introductory information for the site and navigational links to the site's other main pages. Since "home page" has confusing connotations for some readers, we simply call it the site's entry page. In some cases it is possible to use "site entry page" and "site" synonymously. For example, "Visit the CNN site [entry

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA.
Copyright 2001 ACM 1-58113-331-6/01/0009 ...$5.00.

page] at `http://www.cnn.com/`" or "TREC papers may be found on the TREC Web site (`http://trec.nist.gov/`)."

A site finding task is one where the user wants to find a particular site, and their query names the site. A search system succeeds in the task if it returns the entry page of the required site: the "correct answer". If the method returns a list of results, the correct answer should be at, or close to, the top rank. This is different from a subject search, where the user's query describes their topic of interest and the results list should contain as many relevant documents as possible. Site finding is similar to known item search, in that the user is looking for a particular item (site). However, in known item search the user has seen the item before, whereas site finding may involve a known or unknown site. Also, site finding queries name the required site. Known item search queries might describe the topic of an item, rather than naming it.

2.1 Site finding examples

These sentences, taken from an AltaVista log, seem to be site finding queries (although it is impossible to confirm this without user consultation):

> Where can I find Hotmail?
>
> Where is the official Michael Schumacher home page?
>
> Where can I find the web site for Toshiba?
>
> Where is the fun site dating patterns analyzer?
>
> Where is the official Star Wars site?

Notice that all questions start with "where". The user knows which site they want, but not its location (URL).

Taken from the same log, these are probably not site finding queries (but again this categorization is unconfirmed):

> How does a modem work?
>
> What should I consider when purchasing a PC for under $2,000?
>
> What is mp3 and where can i learn more about it?
>
> Who was Cleopatra?
>
> Why do dogs have wet noses?
>
> Where is the Taj Mahal?

TREC systems (including at least one based on the Okapi BM25 formula [10], used in the present study) were found to be highly effective in answering queries of this type, even in comparison to commercial search engines [6].

The AltaVista examples indicate that different user needs exist. "Where is the CNN home page?" is clearly different from "What is CNN?". However, typical Web queries are much shorter: "cnn". Since it is interesting to see which methods work in a typical Web setting, we deal with shorter queries here, such as those listed with "correct answers" in Table 1.

This also raises the question of query disambiguation. It seems impossible to determine whether the user is looking for a specific Web site given a one word query. However, in experiments here we assume the user has knowingly typed their query into a specialist site finding service, hoping to find a specific site.

2.2 Definitional details

We consider site finding queries where the user is looking for a single Web site (see Table 1). The slightly different problem, where the user is looking for multiple sites in a general topic area, is left for a future study. A search for "Bob's U2 Site" would be within our scope, but a search for "U2 Sites" would not.

A site entry page may have multiple equivalent URLs. In such cases, all equivalent URLs are good answers. This occurs in, for example: official mirror sites `http://www.samba.org/` and `http://us1.samba.org/`; alternate domain names `http://www.microsoft.com/` and `http://microsoft.com/`; alternate file names `http://www.yahoo.com/` and `http://www.yahoo.com/index.html`; and pages which forward users directly to another page. This last case includes the multimedia "splash pages", which sometimes precede a site's home page. In all such cases, both URLs are flagged as correct answers during our manual examination of pooled results.

We do not consider an unofficial mirror to be an equivalent answer, since it lacks the authority of the true site. It may be altered or become out of date with respect to the official site.

Evaluation is based on the rank of the first (equivalent) correct answer. No credit is given for returning a page which links to a correct answer. Such pages require work from the user, in finding the page and link. In future it would be possible to take linking pages into account, at the cost of making the experiment more complex. Is a linking page at rank one better than a correct answer at rank ten? Is a correct answer and linking page better than just a correct answer?

2.3 Site finding in practice

A specialist site finding service already exists on the Web: RealNames (`http://www.realnames.com/`). However, it is a hand-made database which maps keywords to entry pages, whereas we consider automatic methods. A number of other systems do not state their purpose as site finding, but seem tuned to do so. Examples are the automatic engine Google `http://www.google.com/` (whose "I'm feeling lucky" button seems geared towards going to a specific Web site) and the hand made directory Yahoo! `http://www.yahoo.com/`. The experimental methodology in this study would allow systems like Google to be evaluated and tuned. It could also compare them to manual systems like RealNames and Yahoo!.

The current evaluation is not the first to consider site finding. However, it is the first published study conducted in a controlled environment with a large number of queries. A previous small experiment by Brin and Page [3], showed that only 1 out of the 4 top commercial engines was capable of finding its own site entry page. Search Engine Watch in March 1998 included a table of whether nine engines could find each other, and has since released three further experiments (detailed at the Web site [12]).

3. BACKGROUND ON LINK METHODS

This section provides background information on link methods and related issues. In our experiments we evaluate a link method based on anchor texts, described fully in Section 5.

Context	Query	Correct answer
VLC2 random	Kennedy Space Center	`http://www.ksc.nasa.gov/`
	AustraLearn	`http://www.clemens.com/AustraLearn/Default.htm`
	Dayspa	`http://www.dayspa.com/`
	Dayton Wide Web	`http://www.daytonwideweb.com/`
	Doug Scoular	`http://ness.arch.su.edu.au/~doug/`
	Metafor Imaging	`http://metaforimaging.com/`
	Geological Data Services	`http://gdsdatamaps.com/`
	Colville Ltd	`http://www.colv.com/index.html`
	Lynnford Leather and Spice	`http://www.lynnford.com/index.htm`
	infoseek	`http://proxy2-bbn.infoseek.com/`
VLC2 Yahoo!	Solomon's Saloon	`http://localsonly.wilmington.net/~solomon/west.html`
	CAKE 2 GO	`http://cake2go.com/`
	INSS	`http://www.usafa.af.mil/inss/inss.htm`
	European Institute for the Media	`http://eim.org/`
	Scottish Liberal Democrats	`http://www.scotlibdems.org.uk/index.htm`
	RAM Mobile Data Limited	`http://www.ram.co.uk/`
	Scope Resume Bank	`http://www.scopeinc.com/resubank.html`
	Prince William Ice Forum	`http://www.pwcweb.com/iceskate/`
	UCLA Alumni	`http://www.bruin.ucla.edu/alumni/`
	VISITING BRITAIN	`http://www.soon.org.uk/britain.htm`
ANU	Strategic and Defence Studies Centre	`http://sdsc.anu.edu.au/`
	Graduate Program in Statistics	`http://www.anu.edu.au/graduate/programs/s3/`
	Wood Workshop	`http://www.anu.edu.au/ITA/CSA/wood.html`
	Centre for Science and Engineering of Materials	`http://www.anu.edu.au/CSEM/`
	Gender Relations Centre	`http://rspas.anu.edu.au/grp/`
	Graduate Program in Political Science	`http://www.anu.edu.au/graduate/programs/p4/`
	Faculty of Economics and Commerce	`http://ecocomm.anu.edu.au/`
	Bruce Hall	`http://brucehall.anu.edu.au/`
	Aboriginal Studies	`http://arts.anu.edu.au/AASchool/abriginalstudies.htm`
	National Centre for Epidemiology and Population Health	`http://nceph.anu.edu.au/`

Table 1: These query pairs were chosen at random from our three test sets. The site finding task was, given a query, to find the correct answer or an equivalent URL. Each correct answer is only one in a set of equivalent URLs. For example, `http://metaforimaging.com/` was judged as equivalent to `http://metaphorimaging.com/`.

3.1 Link-based ranking methods

A hyperlink is a relationship between two documents or two parts of the same document. The source document is the one containing the link. On the Web, the source document would contain text such as:

```
<A HREF="http://www.acm.org/">ACM Site</A>
```

The target document is the one referred to by the link, in this case `http://www.acm.org/`. The link's anchor appears to the user in the source page, in this case with underlined text: ACM Site. If the user selects the anchor, their browser will display the target document.

A ranking method, given a query and a set of documents, generates a ranked list of documents. In the site-finding task, the entry page of the described site should appear as close as possible to the top of the list.

Link-based ranking methods use link sources, targets and possibly anchors to generate the ranked list. Link methods can be divided into three classes, depending on which of these alternate assumptions they rely: recommendation, topic locality and anchor description. The anchor text method evaluated in this paper is based on the anchor description assumption.

The **recommendation** assumption is that by linking to a target, a page author is recommending it. Accordingly, a page with high in-degree is highly recommended, and should be ranked more highly. This can be based on a simple link count [4] or on a calculation of page weights by iterative propagation [3, 8]. A recent study found that link recommendation methods do a good job of picking high quality items, as judged by human experts [1]. Judgments of quality may be necessary to properly evaluate recommendation methods, such as those used by Bharat and Henzinger [2]. However, reliably defining and measuring different aspects of quality, and assessing their importance to end users, is a new endeavor [13].

The **topic locality** assumption is that pages connected by links are more likely to be about the same topic than those which are not. A recent study [5] found that this assumption is often true. The topic locality assumption makes it reasonable to propagate match scores along links via spreading activation or to perform probabilistic argumentation [11].

Using such methods, a page which is link-adjacent to likely-relevant pages may be ranked more highly.

The **anchor description** assumption is that the anchor text of a link describes its target. Using the example link mentioned previously, the anchor text "ACM Site" is describing `http://www.acm.org/`. This makes it reasonable to index anchor text as though it occurred in the target document, rather than (or as well as) the source [9, 3]. The link method evaluated here propagates anchor text in this way (see Section 5).

The ranking methods used by many commercial search engines, including Google [3], incorporate link recommendation information [12]. Google also associates link anchor texts with target documents [3]. We are not aware of any search engines which rank based on topic locality.

3.2 TREC experiments with links

The Text Retrieval Conference (TREC) has primarily concentrated on subject searches, performed over news and government documents. It has recently expanded to new search tasks, such as question answering, and new document sets, including several Web collections [7]. The Web collections are based on a 1997 Internet Archive (`http://www.archive.org`) crawl of over 50 million pages. The 100 gigabyte VLC2 collection is an 18.5 million document subset. The WT2g and WT10g collections are 0.25 and 1.25 million page subsets respectively.

The TREC tasks comparing link and non-link methods were based on a subject search task. The TREC-8 Small Web Track compared many methods over WT2g, finding that links did not help. The TREC-8 Large Web Track incorporation of PageRank [3] in ranking over VLC2 did not enhance effectiveness. The TREC-9 Main Web Task results are not yet finalized. However, indications are that links again did not help, in subject searches over WT10g.

The VLC2 has been criticized on the grounds that its link graph is too small and incomplete for link methods to work. It has only 18.5 million pages, while some commercial engines index over 500 million [12]. However, in this study the link graphs for VLC2 and for a smaller more recent crawl are sufficient.

4. EVALUATION METHODOLOGY

Our evaluation has five steps:

1. **Choose corpus** Choose a fixed test corpus of hypertext documents.

2. **Identify query pairs** Identify a set of <query, site> pairs (for example <sigir 2001, `http://www.sigir2001.org/`>), numbering perhaps 100. Each represents a user typing a query, in order to find a particular site entry page.

3. **Run methods** For each method being evaluated, run the queries over the corpus.

4. **Examine results** For each query, examine pooled results to identify equivalent URLs (e.g. mirror sites).

5. **Measure effectiveness** Apply some effectiveness measure. In case of multiple equivalent correct answers, measure according to the top ranked one.

The test corpus should simply be a document set where effective site finding is important, such as the general Web crawl and university crawl used here. Each query pair describes the user's query and underlying need. They type the query "sigir 2001" in order to find the SIGIR 2001 Web site at `http://www.sigir2001.org/`.

For each query, ranking methods are applied. In this case we limit the output to the top 10 ranked documents. This is because most Web search engines produce a top 10 ranking by default, so the most desirable place to find a correct answer is (near the top of) the top 10.

Pooled results examination may seem similar to that done in TREC ad hoc experiments, presenting the union of systems' results for human judging. However, the judging criteria are very different. In TREC, judges identify documents in the pool that would be relevant to a querying user. Here, judging simply identifies other URLs which are equivalent to the expected correct answer in ways described in Section 2.2. This is required for fairness, so that each of the equivalent answers is rewarded.

The most commonly quoted effectiveness measure in TREC is average precision. If only the top ranked correct answer is counted, average precision is equivalent to the reciprocal rank of the top ranked document. Accordingly, we chose mean reciprocal rank, within the top 10, as our principal measure. We also present precision at one document retrieved (P@1), a histogram of ranks at which the correct answer appeared and a pairwise comparison of methods (number of queries where A was superior and number of queries where B was superior).

The most difficult aspect in experimental design was deciding how to choose the query pairs. Pairs chosen should have the following properties:

- Spread of sites: If a certain population of sites are of interest, a representative spread of sites should be included in the evaluation. For example, an experiment would be of limited impact if it only considered sites which are of interest to computer scientists.

- Realistic query formulation: Web users typically type a short query, which may not be the precise name of the site required. The site finding queries used in evaluation should have these qualities, perhaps containing abbreviations, missing words from the site's official name or adding extra words to it.

Ideally we would have used a large sample of query pairs, from a representative sample of Web users. Unfortunately, there is no publicly available test collection of this type and we lacked time and resources to undertake a methodologically sound survey of Web users.

Falling short of this ideal scenario, we considered two approximative methods for preparing the data. We refer to them as query-first and document-first selection of <query, site> pairs.

Query-first selection involves randomly choosing queries from a search log and trying to identify a site that is a plausible correct answer for each query. For example, choosing a query "sears" from the search log, the experimenter might identify `http://www.sears.com/` as the correct answer. The general problem with the query-first approach is that the experimenter's interpretation of what is a "plausible correct answer" might not reflect the real user need underlying each query.

We encountered an additional problem with the query-first approach in the context of our VLC2 experiments. Due to the incompleteness and age of the VLC2 data (it is one third of a 1997 crawl) only a small proportion of queries have a correct answer in VLC2. Queries with non-VLC2 answers should be excluded from this experiment, to give the content-based method some chance of succeeding. (Link methods are capable of finding non-VLC2 sites, using VLC2 link information.)

Document-first selection, on the other hand, involves randomly picking site entry pages and manually generating an appropriate query. Random entry pages can be selected, for example, from Web directories. Alternatively, one can choose pages randomly from the collection and navigate to the entry page of the site that contains the page, either by traversing links or by editing the URL (if no suitable links are available). For example, if the randomly chosen page is a registration page for the KDD'01 Conference, navigation would lead to the conference site entry page.

The document-first approach has the advantage of generating a representative sample of the crawl's population of sites. Its disadvantage is that queries formulated by experimenters might not be representative of queries submitted by the general user population.

Since, in practice, it is easier to deal with the query generation problem we decided to apply both variants of the document-first selection method.

5. EXPERIMENTS

We compare two simple methods, a link anchor method and a content method, with minimal differences. Both methods use the same implementation of the Okapi BM25(2.0, 0.0, inf, 0.75) formula from [10, pages 110–111], including length normalization controlled by $b = 0.75$. In neither case is relevance feedback or stemming employed.

The content method indexes the textual content of each document. For the query "excite" the correct VLC2 answer document is `http://www.excite.com/` whose textual content is:

> Excite Home World's best news, FREE! Chocolates & Wine Cards & Music Flowers Gifts Excite Search Twice the power of the competition. Search the entire Web Search NewsTracker Search Excite Web Reviews Search Usenet newsgroups Search Tips Advanced Search Submit For info on destinations around the globe, visit City.net. Reference People Finder Yellow Pages Email Lookup Travel Search Shareware Resources Free Start Page Bookmark Excite Excite Direct New to the Net? Free Search Engine Information Help Feedback Advertising Add URL About Excite Jobs at Excite Excite Web Reviews Our insights into the . . . [106 more words]

This is ranked, using BM25, against the other 18.5 million documents.

The anchor method does not use any of this text. Instead, based on the anchor description assumption, it builds "anchor documents". Each anchor document contains all the anchor texts of a page's incoming links. If 7 332 pages link to `http://www.excite.com/` with the anchor text "excite", that word is added 7 332 times to the anchor document. Based on VLC2 links, the correct answer's anchor document contains:

> 7 332 × excite
> 910 × excite netsearch
> 294 × http://www.excite.com/
> 227 × excite search
> 200 × excite!
> 192 × http://www.excite.com
> 168 × e xcite
> 154 × view
> 140 × excite home
> 86 × excite search engine
> 66 × excite search:
> 49 × exite
> 42 × www.excite.com
> 35 × (www.excite.com)
> 28 × excite:
> 28 × [excite]
> 23 × *excite
> 21 × e x cite
> 18 × excite net search
> 17 × o ptions
> . . . [440 more lines]

There are 11 000 links from VLC2 pages to the Excite entry page, each on average containing 1.3 words, so the anchor document is quite large. It is ranked, again using BM25, against 44.1 million anchor documents.

The two methods are first compared using the VLC2 documents and two different sets of query pairs. Then we use a recent 400 000 document crawl of Australian National University Web servers.

The first set of 100 pairs was generated by choosing a random VLC2 page and navigating to the nearest entry page, then generating an appropriate query. In about one third of cases navigation was not possible, due to the entry page being outside VLC2 or the page being non-English (making later query generation too difficult). These pages were discarded.

The second set of 100 pairs was generated using Yahoo! listings in the VLC2 collection. If the randomly chosen Yahoo! link was live in VLC2, we generated an appropriate query. Since Yahoo! is a selective manually maintained Web directory, listed sites may be more popular or well known than the average Web site.

A third set of 100 pairs was generated based on a page listing hundreds of Web sites within the university. Most sites corresponded to a teaching or research unit within the university. If a randomly chosen page was within the university crawl, we generated an appropriate query.

For each set of sites, queries were generated by the experimenters. Queries were based on the site's name, usually without spelling mistakes. In some cases names were replaced with abbreviations. In other cases words were added or omitted. This ensures that there is no simple pattern, such as the query always precisely matching the title of the page in question. Although we endeavored to keep queries short, we did not sacrifice preciseness to do so. The average query lengths (including stop words) for the three sets of queries are: 2.9, 2.7 and 3.3 words respectively. Example pairs are in Table 1.

Chosen query words almost always occurred in the required document. This was simply because most company Web sites contain the company name, and personal sites contain the person's name. If anything, this occurrence of query terms in document full texts should favor the content-based method.

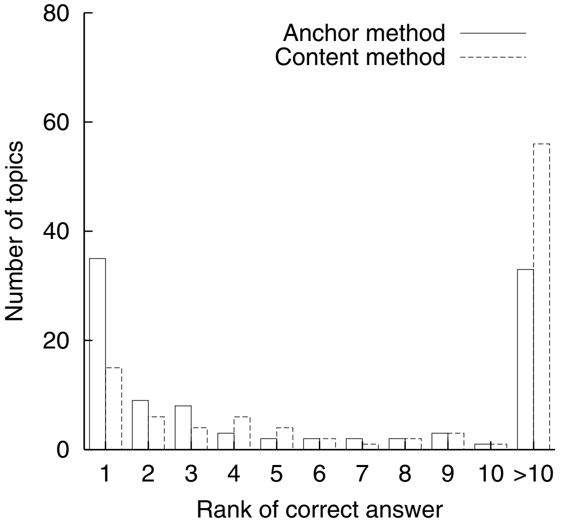

Figure 1: VLC2 results for 100 site entry pages, chosen randomly through page selection and navigation. For 35 of the 100 queries, the anchor method returned the correct answer at rank one, compared to 15 times for the content method.

5.1 Results

For the 100 randomly chosen VLC2 sites, the mean reciprocal rank of the first correct answer within the top 10 was 0.228 for the content method and 0.446 for the anchor method. The distribution of correct answer ranks is depicted in Figure 1. P@1 was 0.15 for content and 0.35 for anchors. Anchors achieved a better rank on 55 queries, equal on 24 queries and worse on 21.

For the 100 Yahoo!-listed sites, the content method scored a mean reciprocal rank of 0.370. The anchor method achieved 0.720. The distribution of correct answer ranks is depicted in Figure 2. P@1 was 0.27 for content and 0.62 for anchors. Anchors were better on 59 queries, equal on 26 queries and worse on 15.

For the 100 university sites, the content method scored a mean reciprocal rank of 0.321. The anchor method achieved 0.790. The distribution of correct answer ranks is depicted in Figure 3. P@1 was 0.21 for content and 0.68 for anchors. Anchors were better on 72 queries, equal on 24 and worse on 4.

In each of the three cases, a sign test showed that the differences were significant ($p < 0.01$).

6. DISCUSSION

The most important feature of the results is the strong advantage of the anchor method over the content method. Since both used the same retrieval system and ranking algorithm, the best explanation for this is that anchor information is more useful than content on this site finding task. The strength of the difference indicates that anchor texts are highly useful in site finding. It also uncovers conditions under which well known content based methods, which are consistently effective in TREC subject search tasks, are less effective.

The experiments were not biased towards the anchor method. Queries were generated after viewing the document content,

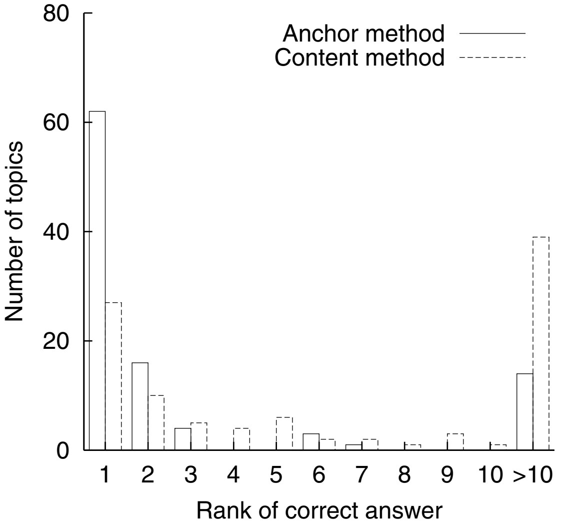

Figure 2: VLC2 results for 100 Yahoo!-listed sites.

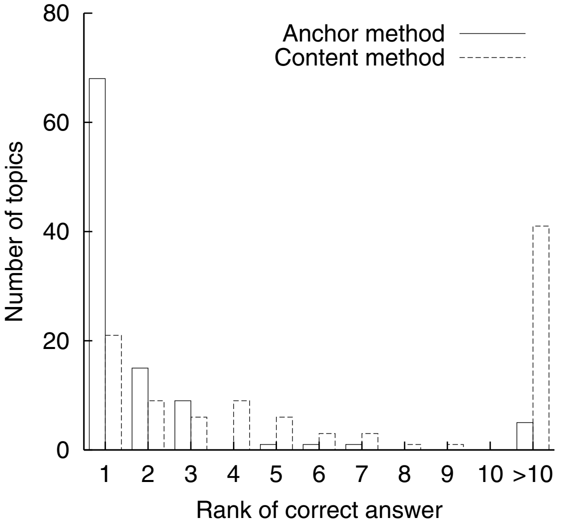

Figure 3: University results for 100 sites within the institution.

and as observed in Section 5, query terms almost always occurred in that document. By contrast, it was not checked whether query terms occurred in the anchor document, or even whether a non-empty anchor document existed. Empty anchor documents occur when a page has links with image anchors rather than anchor text.

In addition to the query selection, correct answers were restricted to the test collection, to ensure that the content method had some chance of finding them. During query selection many sites retrievable by the link method were rejected for this reason.

Another feature of the results is that, in Figures 1–3, the biggest difference between link and content methods is at rank one. Considering only ranks 2–10, there is very little to separate the methods. This raises the question as to whether the correct answers at rank one are different from the other answers.

In the first experiment, the anchor method returned 35 correct answers at rank one. Of the 35 results the median in-degree was 35. For the full set of 100 answers it was 21. Perhaps unsurprisingly, the anchor method tended to succeed on answers with more incoming links. The exception were two failures with very large in-degrees: Excite with around 11 000 links and Yahoo with 115 000. In each case, given queries "excite" and "yahoo", the correct answer was not in the top 10. Perhaps with a very large number of links, some "pollution" occurs, introducing noise words to the anchor document which have the overall effect of lowering its BM25 rank.

For one third of the randomly chosen sites, the anchor method did not return the correct answer in the top ten. Improving this would be desirable. However, the link method employed here is very simple. It was chosen because it is an absolutely pure link method, using no information other than the link anchor and target. Superiority of a pure link method against a pure content method provides strong evidence that anchors are useful for site finding. There are also no confounding factors such as use of differing software.

In future experiments it will be interesting to test other link and content methods, and combinations of methods. It would also be possible to test well known retrieval techniques such as relevance feedback, to measure their impact on site finding. Methods may well be found which would reduce the number of sites outside the top 10.

The second VLC2 experiment (Figure 2) showed results similar to the first (Figure 1) but with higher absolute values. This may be because Yahoo!-listed sites are more heavily promoted by their publishers. Site promotion may generate more incoming links (helping the anchor method) and involve seeding the entry page with appropriate search keywords (helping the content method). However, for fair accessibility of information, automatic methods should also be effective for the general (randomly selected) sites.

The link anchor method was highly effective in the third experiment. This was achieved despite the small size of the university crawl. This is almost certainly because of many hyperlinks between different teaching and research sites within the university.

At this stage no conclusion can be drawn as to whether a retrieval system should provide distinct site finding and subject search subsystems, or whether a composite ranking function can be found which excels at both tasks.

One way of looking at incoming links is to see the anchor text as (part of) a "query" to which the linked document is an "answer". Thus the <anchor text, linked document> pair constitutes a sort of training query. This view opens up several possibilities which may be explored in the future, including combining anchor text evidence with other sources of training "queries".

This study shows that anchor texts can be useful when the query is the name of a site. Anchor propagation also has other interesting properties. It can succeed when the required page is outside the crawl or contains no text. It successfully found `http://www.csiro.au/` given the query "csiro" using the VLC2 and university anchor indexes, even though neither crawl contains the page. It can even enable retrieval with common misspellings or alternate names. In our VLC2 anchor index "quantas" found `http://www.qantas.com.au/` and "uk prime minister", "number ten", "british prime minister" and "downing street" all correctly returned `http://www.number-10.gov.uk/`.

A possible criticism of these experiments is that the content method is not ideal. BM25 could become more effective when employed with stemming, relevance feedback and passage scoring. Alternatively a different tuning of BM25 parameters or completely different content method might be even more effective. We certainly do not discount these possibilities. However, we have unambiguously shown that a well known and consistently effective subject search method is not ideal for this site-finding task. We have further shown that even an untuned link anchor method is capable of highly effective site finding. (It ranked the correct answer within the top 10 results for 95% of the university site finding queries.)

In terms of search efficiency, the VLC2 content index covers 18.5 million documents with a total of 5 669 million term occurrences. The anchor index covers 44.1 million link targets, with a total of 411 million term occurrences. Although the link index contains more documents (including many documents outside VLC2) it contains many fewer term occurrences, since only anchor texts are included. This means that a site finding index is potentially an order of magnitude smaller than a full text index. However, gains from this might be partially offset by having to deal with more documents at query time. In addition, the anchor index is quite specialized, and might not be useful in for other types of query.

A final consideration is "spam", where page authors deliberately manipulate their links or content to boost their page rankings. We took no anti-spam measures in our experiments. Thus, our ranking methods could have been fooled by repetition of our query words in content or anchor text. However, spam tends to target popular queries, whereas we used less popular queries (see Table 1). Also, our results were confirmed in a crawl of the Australian National University (Figure 3), an environment where spam is not a widespread problem.

7. CONCLUSIONS AND FURTHER WORK

Anecdotal evidence suggests that site finding is a common and important Web searching activity. In each of three site finding experiments, we found that BM25 ranking applied to link anchor documents significantly outperformed the same ranking method applied to document full texts. In each case and for both of the measures employed, the link anchor method was approximately twice as effective as the content method. The anchor method worked equally well on 18.5 million document and 0.4 million document collections.

We conclude that link anchor text provides a valuable source of information for use in site finding tasks, even within collections less than 0.1% of the size of the Web.

Commercial search engines have been using a variety of link-based ranking methods for a considerable time. However, there have been no published evaluations of the performance of such methods on site-finding tasks. Using the methodology of the present study as a base there are a great many aspects of this important problem to be investigated in future work. Can specialized full-text (non-link) methods be found which achieve similar performance to the link anchor method? How much can the performance of the link anchor method be improved by tuning? How well do other link methods work? Are there any other forms of descriptive text, besides link anchor text, which can be used in site finding? What is the optimal weighting of the various forms of link and content evidence on site finding tasks? How useful is anchor text evidence on other search tasks?

8. ACKNOWLEDGMENTS

Thanks to Natasa Milic-Frayling for her very helpful comments. Thanks to Steve Walker for technical advice at Microsoft Research.

9. REFERENCES

[1] Brian Amento, Loren G. Terveen, and William C. Hill. Does “authority” mean quality? Predicting expert quality ratings of Web documents. In *Proceedings of SIGIR 2000*, pages 296–303. ACM, 2000.

[2] Krishna Bharat and Monika R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of SIGIR'98*, pages 104–111, 1998.

[3] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of WWW7*, 1998.

[4] S. Jeromy Carrière and Rick Kazman. WebQuery: Searching and visualizing the Web through connectivity. In *Proceedings of WWW6*, pages 701–711, 1997.

[5] Brian D. Davison. Topical locality in the Web. In *Proceedings of SIGIR 2000*, pages 272–279. ACM, 2000.

[6] David Hawking, Nick Craswell, Paul Thistlewaite, and Donna Harman. Results and challenges in Web search evaluation. In *Proceedings of WWW8*, pages 243–252, 1999.

[7] David Hawking, Ellen Voorhees, Peter Bailey, and Nick Craswell. Overview of TREC-8 Web Track. In *Proceedings of TREC-8*, pages 131–150, Gaithersburg MD, November 1999. NIST special publication 500-246, `http://trec.nist.gov`.

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

[9] Oliver A. McBryan. GENVL and WWWW: Tools for taming the Web. In *Proceedings of WWW1*, 1994.

[10] S. E. Robertson, S. Walker, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of TREC-3*, pages 109–126, Gaithersburg MD, November 1994. NIST special publication 500-225.

[11] Jacques Savoy and Justin Picard. Retrieval effectiveness on the Web. *Information Processing and Management*, 2001. To appear.

[12] Danny Sullivan. Search engine watch. Web Site, 2001. `http://searchenginewatch.com/`.

[13] Xiaolan Zhu and Susan Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. In *Proceedings of SIGIR 2000*, pages 288–295. ACM, 2000.