How Valuable is External Link Evidence when Searching Enterprise Webs?

David Hawking[1] **Francis Crimmins**[1] **Nick Craswell**[1] **Trystan Upstill**[2]

1. CSIRO Mathematical and Information Sciences,
GPO Box 664,
Canberra, Australia 2601
David.Hawking@csiro.au
2. Computer Science Department, ANU,
Canberra Australia 0200

Abstract

Link information, especially anchor text, is known to be very useful for effective ranking of web pages, particularly in response to navigational queries. We investigated whether enterprise webs contain sufficient internal link information to adequately answer queries derived from the enterprise's site map or, alternatively, whether adding link evidence from the external Web can boost search effectiveness. Using 1266 navigational queries derived from Stanford University's A-Z site index, we found no difference between the quality of results returned by Stanford's Google appliance and those from an appropriately site-restricted search of the global Google service. Applying similar methodology to our own crawls of seven Australian organisations, we found that adding external link evidence made no significant difference to search effectiveness in five cases and a slight difference (in different directions) in the other two. We observed that external links to an organisation show very different patterns to internal links. Unlike enterprise web publishers, external web authors heavily favour directory default pages, particularly the organisation's home page and pages offering information or services likely to be useful on an ongoing basis. External links seldom reference the complex, parameterised URLs in common use in many organisations.

Keywords: Enterprise search; search quality evaluation; hyperlink methods.

1 Introduction

Many organisations, be they commercial or not-for-profit, invest heavily in publishing information and services to the external world via the public Web. The likely return on this investment depends upon how easily potential "customers" can locate the information and services using, both:

1. Public whole-of-Web search services such as `www.google.com`, `www.alltheweb.com`, and `search.msn.com`; and
2. The enterprise's own site search facility, which is normally restricted to cover only the internet domain(s) controlled by the enterprise. For example a search of `nist.gov` operated by the National Institute of Standards and Technology (NIST), or a search of `ibm.com` operated by IBM.

When ranking answers to a query, whole-of-web search engines prefer web sites (represented by their home page) to individual web pages and highly recommended pages (those with many incoming links from trustworthy sources) ahead of less recommended ones. (See Section 1.1 below for explanation of terms such as page and web site.) A web site is generally a more useful answer to a query than an individual page because it represents a structured package of relevant information, from a usually identifiable source, within which the visitor may browse or search. The homepage of a web site generally provides an overview of the information available and indicate the structure of the site.

For example, the query `Sony` will return the home page of the Sony Corporation as the first result on most whole-of-Web search engines. Once at `www.sony.com` the searcher can choose whether to visit the Sony shop, to log in to the Sony members site, to find out more about Sony electronic products, Sony online games, Sony music etc, or to search within the site. By contrast, although the page `www.sonystyle.com/is-bin/INTERSHOP.enfinity/eCS/Store/en/-/USD/SY_BrowseCatalog-Start?CategoryName=acc_MemoryStickMedia&Dept=acc` is both relevant to the query and authoritative, it is too specific (covering only the topic of Sony memory sticks) and doesn't give a good overview of the general topic.

Similar principles apply in web searches restricted to a single organisation's web. Within such a web (e.g. the Sony web represented by the domains `sony.com`, `sonystyle.com`, etc.) similar but smaller sites are observable corresponding to products (eg. `station.sony.com`) divisions (`us.sonymusic.com`) and activities (eg. `www.my.sony.com`). Quality search services at the organisational level will also return relevant sites ahead of relevant individual pages. Anyone who has operated the organisational search service for a university will know that clients are very unhappy if the first response to the query `library` is not the homepage for the main library web site. Similarly, most searchers at `www.microsoft.com` would expect to find the home page for the MSWord product when they type the query `word`.

It is understood that whole-of-Web search engines make extensive use of link anchor text and measures based on hyperlink graph structure, such as PageRank. (Brin and Page, 1998) Unfortunately, an enter-

prise search engine generally indexes only pages published by the organisation and cannot see potentially large quantities of useful link information derivable from the Web as a whole. Naturally, not all of this external evidence will be beneficial, but it seems plausible that a search engine with access to external link information (particularly a large crawl of the World Wide Web) might deliver better responses to navigational queries than one operating entirely within the enterprise context.

Accordingly, this paper will attempt to test the following null hypothesis:

> Null Hypothesis: *A search engine with access to external link information (particularly a large crawl of the World Wide Web) will in general perform no better on navigational queries than the same engine operating entirely within the organisational context.*

Notes:

1. Navigational queries are those where the searcher types the name of some entity as a query and wants the search engine to find the home page corresponding to that entity. (Broder, 2002, Craswell et al., 2001) For example the query `Sony` on the Web or the query `library` to a university search engine.

2. The null hypothesis is self-evidently true for engines which take no account of link information.

We hope to arrive at a conclusion with a useful degree of generality by using search engines known to make extensive use of link information and by choosing, within the limits of available information, a diverse set of organisations for study.

Even ballpark guidance on the central question is likely to be of practical value to organisational webmasters and CIOs faced with a choice between deploying their own search engine and presenting scoped results from a whole-of-Web search engine like Google.

The first experiment presented here is a comparison of homepage finding effectiveness between `www.google.com` with a `site:stanford.edu` scope restriction and the Google Appliance operated by Stanford University (`find.stanford.edu`).

Although the outcome of this comparison is potentially interesting, we have no guarantee that the algorithms employed by the two different Google products would operate at equal effectiveness on the same data. Accordingly, we present some additional studies where the ranking algorithm can be held constant (or varied as required).

We studied seven Australian organisations chosen to represent a range of web site sizes and a diversity of organisational types (three commercial, two educational and two government). Our crawls of these organisations ranged from 2,000 to 170,000 pages. We crawled each of the organisations in question and produced two indexes for each, one using local evidence only and the other using all the incoming links from a six-million page subset of the Australian web (`.com.au`, `.gov.au` and `.edu.au`). This subset represents a reasonable fraction of the full `.au` domain, Google[1] indexes approximately 6,980,000 `.au` pages.

[1] accessed on 28 August 2003.

We did not have access to a whole-of-Web crawl but hypothesised that a substantial proportion of incoming links to the chosen Australian organisations would originate within Australia.

The Panoptic search engine[2] was used in these studies – it is specifically designed to address navigational as well as other search tasks and makes use of both anchor text and inlink evidence.

Please note that Panoptic is sold commercially and that three of the present authors are members of the Panoptic team. We have explicitly avoided making claims for the product or comparisons with competitors.

Section 2 reviews past work while Section 3 discusses the evaluation methodology. The main experiments and their results are presented in Sections 4 - 6 and discussed in Section 7.

1.1 Terminology

anchor text	The highlighted words people click on in their browser in order to follow a link.
enterprise	Any commercial, research, educational, not-for-profit or government organisation controlling an internet domain.
home page	The web page intended as the main entry to a web site.
site map	A list maintained by the web site owner of key web pages or subsidiary sites within the web site.
URL	Uniform Resource Locator - the address of a web page. e.g. `http://www.xyz.com/About/contacts.html`
web page	A document addressable by a URL and deliverable by means of the HyperText Transfer Protocol (HTTP). Each is a node in the web connectivity graph.
web site	A coherent, interdependent collection of web pages published by a single publisher and representing an entity such as a person, an organisation or a product. Each site has a home page. Note that sites may contain subsidiary sites. e.g. The Sony site `sony.com` contains subsidiary sites representing both products and divisions of the company.

2 Related Work

Broder (2002) highlights the importance of trying to address the need behind Web search queries and characterises Web search loads as a mixture of Informational (the intent is to acquire some information assumed to be present on one or more web pages), Transactional (the intent is to perform some web-mediated activity) and Navigational (the immediate intent is to reach a particular web site) searches. Singhal and Kaszkiel (2001) show that retrieval methods designed for informational tasks do not work well on representative Web search. Navigational search has been studied extensively within the TREC Web Track. (Hawking and Craswell, 2001)

Many authors have addressed the problem of presenting a high quality result set when there are very large numbers of documents which fully match the query. Kleinberg (1997) proposes link methods for identifying *authoritative sources* starting with a list of search results. Chakrabarti et al. (1998) use similar

[2] Panoptic version 5.0.0.21, `www.panopticsearch.com`

methods for *resource compilation* while Bharat and Henzinger (1998) refine Kleinberg's method and describe the problem as *topic distillation*. Topic distillation has recently become a TREC Web Track task. (Craswell and Hawking, 2002)

Calado et al. (2003) discuss global and local link graph measures but, in their context, local refers to a small query-specific link graph within a large web and does not refer to organisational web sites.

Use of topic-independent link graph metrics such as inlink count Carriere and Kazman (1997) and PageRank (Brin and Page, 1998) to identify popular or important pages are well known and have been investigated within enterprise-scale collections by Upstill et al. (2003).

The use of anchor text in Web search has been reported by McBryan (1994) and Brin and Page (1998) while Davison (2000) and Amitay and Paris (2000) have explored the use of hyperlinked descriptions beyond the explicit link anchor. The usefulness of anchor text in enterprise-scale collections has been confirmed by Craswell et al. (2001).

Bharat et al. (2001) observe that "the Web is actually a hierarchically nested graph with domains, hosts and web sites introducing intermediate levels of affiliation and administrative control" and study paterns of linking between hosts. We have been unable to locate prior work on the types of pages which tend to be the targets of links.

3 Experimental methodology

This section explains how queries and corresponding right answers were obtained for the organisations studied, how queries were submitted and results collected, what measures were used, how the issue of equivalent right answers was dealt with, and how two sets of results for the same organisation were compared.

3.1 Obtaining queries and right answers

Many organisational web sites include a site map or site index containing a list of what the organisation considers to be the important sites within its web. Such site maps were used as low-cost sources of navigational queries and answers.

The site map contains a list of site names together with links to the sites themselves. Such lists (downloaded and saved from the Web) were converted semi-automatically into a set of queries and their corresponding preferred answers. If the site map consisted of multiple pages, the individual HTML files were concatenated. Site maps were edited to remove headers, footers and navbars and then processed using the `make_queries.pl` script, which:

1. Eliminated items for which the specified answer lay outside the domains of interest, included a protocol specification other than `http://`, or lay in a `cgi-bin` directory.
2. Removed HTML tags such as `<font>`,
3. Replaced '@' with ' at ',
4. Replaced the entity '&' with ' and ' and other entities with spaces,
5. Replaced all characters other than letters, digits, period, hyphen, single-quote and double-quote with a space,
6. Removed all superfluous spaces.

In order to facilitate matching, URLs were canonicalised in both `make_queries.pl` and in the scripts which scanned result list rankings using the same function which:

1. Stripped the protocol prefix if present,
2. Converted the URL to lower case[3]
3. Appended a trailing slash if the URL contained none,
4. Removed a `:80` port indication if present, and
5. Stripped off suffixes indicating the default page for a web directory such as `index.html` and `default.asp`.

For example, the A-Z Subject Index page for NIST (`www.nist.gov/public_affairs/siteindex.htm`) would result in a set of query/answer pairs including:

```
Advanced Encryption Standard
 -> csrc.nist.gov/encryption/aes/
Building and Fire Research Laboratory
 -> www.bfrl.nist.gov/
General Publications
 -> www.nist.gov/public_affairs/genpubs.htm
World Trade Centre Investigation
 -> wtc.nist.gov/
```

To serve as a useful test of navigational effectiveness a navigational query/answer set should contain sufficient (preferably hundreds) of queries to average out performance variability across queries.

3.2 Issues with use of the site map in generating queries

A number of general points need to be made about the use of site maps as sources of queries and answers. First, a site map is likely to generate only queries for which there is a suitable single answer within the enterprise web.

Second, the language used in the queries will be that used officially within the organisation. (The Stanford site map is quite comprehensive and includes entries for both 'employment' and 'jobs' but not for 'recruitment' or 'vacancies' or 'positions vacant'.) These first two points suggest that performance on a site map-based test is likely to be higher than for queries using the language and mental models of a range of visitors to the site.

Third, from the point of view of a visitor, the site map may be substantially incomplete: Of the 15 most popular queries submitted to the CSIRO external search engine[4] over a period in early 2003, only one ('employment') was represented in the official site map.

Fourth, search effectiveness on queries for which there is an exact entry in the site map is arguably less

[3] This is not strictly correct, but sometimes avoids missed matches. The probability of seeing a false match from a case sensitive server where the lower case answer is correct but an upper case version is incorrect is considered negligible.

[4] `www.csiro.au`

useful than for queries where there is no such entry. However, we expect that relative performance on site map queries will be predictive of relative performance on other similar queries.

Fifth, the use of site map-derived queries should guarantee that every correct answer will have at least one incoming link and at least one piece of highly descriptive anchor text – from the site map itself. This clearly has the potential to inflate site map-measured effectiveness even further beyond that for arbitrary queries and to diminish the measured value of external evidence. In the Panoptic experiments, we controlled for this by removing the site map page before indexing.

An alternative to use of the sitemap would be to use queries submitted by visitors to the site. However, we only had access to query logs for one of the eight organisations. In any case, if we had used query logs we would have had to try to select those queries with navigational intent and to determine the correct answer if one existed. Neither of those are easy, particularly for someone external to the organisation being studied. Furthermore, a correct answer might be missed, resulting in a possible bias in the evaluation.

Another alternative would be to use referring anchortext on links from outside the organisation. However, this would have biased the queries to those most likely to benefit from the use of external evidence. Indeed, query/answer pairs derived by this means could give credit for returning wrong or superseded answers.

Despite the various issues, enterprise site maps provide a source of navigational queries and judgments whose quality would be hard to match by any other means. In general, site maps are developed to facilitate the business of the enterprise by experts who are intimately familiar with the enterprise and who are in a position to make good judgments about what entries should be included and what are the corresponding best answers.

3.3 Submitting queries and collecting results

An evaluation script `test_effectiveness.pl` (tailored to the particular search engine) was used to submit the queries one by one to the search engine interface via the Web and to look for the first occurrence of a right answer among the search results. The script is written to identify search results proper as opposed to advertising, site navigation, cache links, next and previous result links and other extraneous links. Two general heuristics expressed as regular expression patterns simplify the work:

- Reject any result URL which is not within the domains of interest. This removes offsite advertising.
- Reject any result URL which specifies the search host. This removes cache links, and next and previous page links.

These were augmented as necessary by engine-specific heuristics.

The number of search results extracted served as a test of correct operation. If extraction is working correctly, the number of results retrieved should usually be the default number (typically ten or twenty) returned by the search engine and never more than that number.

The `test_effectiveness.pl` script recorded the result pages for each query (to enable post hoc validation and re-analysis if required) and also generated a list of the queries together with the rank $(1-10)$ at which the first right answer was found. If the search system returned more than 10 answers by default, answers beyond rank ten were ignored. A rank of 11 was recorded if no right answer was returned. Here is an example of a list of queries and the corresponding right-answer ranks:

```
Advanced Encryption Standard -> 1
Building and Fire Research Laboratory -> 3
General Publications -> 11
World Trade Centre Investigation -> 7
```

A script `rundiff.pl` was used to compare two output files generated from the same query set, displaying the queries (and answers) for which there was a difference in ranks. It also reports the numbers of queries for which run A was inferior, superior, and equal to run B and performs a Wilcoxon matched-pairs signed-ranks text. (Klugh, 1970)

3.4 Measures

Script `compute_measures.pl` was used to compute S@1, S@5, S@10 and MRR1 measures from the output of `test_effectiveness.pl`. S@n means success at n documents retrieved and reports the proportion of queries for which a correct answer was retrieved by rank n. MRR1 is the mean reciprocal rank of the first right answer found. (The reciprocal rank is considered to be zero if no correct answer is found by rank ten.) In the four-query example above, S@1 = 0.25, S@5 = 0.5, S@10 = 0.75, and MRR1 = $\frac{1+1/3+1/7}{4}$.

The S@1 measure is most likely to distinguish between different algorithms or systems but is very susceptible to random variation. Users may be expected to be delighted if right answers usually appear at rank one, but in terms of time spent searching, the proportion of queries for which the search engine returns the right answer on the first page of results, "above the fold"[5] is probably the most useful measure. It is most closely approximated by S@5.

From a research point of view, the MRR1 measure is superior because it takes into account all available ranking information. (As far as S@5 is concerned an answer at rank 2 is no different from one at rank 5, but MRR1 scores the two differently.)

3.5 Dealing with equivalent URLs

It is vital to deal in an unbiased way with the issue of unidentified URLs which are equivalent to those listed as correct answers. In a web context it is very common for the same page (or pages with identical content) to be accessable via more than one URL. Some are simple syntactic variants of the same canonical form but other distinct URLs arise from hostname and path aliasing, dynamic page generation and redirection. For example, `www.curtin.edu.au/curtin/dept/appchem/` and `chemistry.curtin.edu.au/` are equivalent.

To obtain a correct evaluation of a result ranking (e.g. $A1, A2, \ldots, A10$), *while remaining true to the judgments of the people who created the site map*, it

[5] i.e. visible without the need to scroll down or navigate to the next page

Figure 1: **Equivalence judging algorithm**: A low-cost, unbiased method for determining whether results lists contain a page equivalent to a right answer above the first occurrence of a previously identified right answer.

1. Canonicalise the URLs in each list.
2. Truncate each list at the first occurence of a known correct answer.
3. Interleave the lists of URLs, randomly choosing the order of interleaving.
4. If the interleaved list is non-empty, present the correct answer to the equivalence judge in a browser window and extract the page title.
5. Process each item in the interleaved list in turn as follows:
   (a) Skip if this URL has already been seen.
   (b) Fetch the page using a text-only browser such as `lynx`[6] which follows redirects and extract the title of the page actually fetched. Skip if the title is different to the one for the known correct answer.
   (c) Present the page for judging in a second (graphical) browser window.
   (d) If judged equivalent, record this URL as an additional right answer and purge the interleaved list of all further items from this result list.

is necessary for the page corresponding to each result URL in the ranking to be compared with the original correct answer, going down the ranking until the first right answer is encountered. Any page which is visually the same as a right answer, and includes the same links, should be accepted as correct.

When comparing runs, it is important that the equivalence judge is not aware of which run contributed which results, while at the same time minimising the number of judgments required. A judging script `equivalence_judging.pl` was developed for the project, which checked a set of rankings $R1, \ldots, Rn$ for each query using the algorithm shown in Figure 1.
Notes:

1. The judgment of which is the best answer for a query remains with the creator of the site map (who is presumed to possess intimate knowledge of the enterprise and its web sites.) Equivalence judging is only about identifying pages which are identical as far as a person viewing them is concerned.
2. No judging is performed if both runs retrieve the right answer at rank one.
3. It might seem that a way of avoiding manual judging altogether would be to download the pages being compared using Lynx or a similar browser and to compare checksums. However, we were concerned that dynamically generated pages might include variable content such as date/time or URL and cause us to miss some correct answers.
4. The augmented answer set resulting from a judging run is still likely to be incomplete but additional answers will reduce the cost of judging subsequent runs.

Table 1: Effectiveness comparison of the `www.google.com` index (site-restricted to `stanford.edu`) against the Google appliance operated by Stanford university, using homepage finding queries derived from Stanford's A-Z site map.

	Queries	P@1	P@5	P@10	MRR1
Appl'nce	1266	.6445	.7322	.7480	.6817
Google	1266	.6216	.7180	.7401	.6639

In experiments comparing two Panoptic search engine runs, it was not necessary to perform equivalence judging because the lists of redirects and duplicate pages encountered during the Panoptic crawl were used to augment the answer set. Thus all identical pages in the collection were already known.

4 Stanford experiments

Stanford University was selected as the site on which to compare the performance of the Google Search Appliance[7] operating at `find.stanford.edu`) with a scoped version of the Google whole-of-Web search service (`www.google.com` with a `site:stanford.edu` restriction). Stanford is a high-profile, single-domain site where the appliance is unlikely to be inappropriately installed or incompetently managed – A press release quoted on the Google corporate web site draws attention to the Stanford installation; and the original research behind Google was conducted at Stanford.

Stanford University (`stanford.edu` domain) presents a site map organised by letter of the alphabet and segmented into 32 separate chunks. These were combined and processed into a set of 1447 query/answer pairs which were run on both search services on 11 May 2003.

The number of `stanford.edu` pages in the main Google index was estimated (using the query `site:stanford.edu -adsjasldjaldjka` on Google) to be around 663,000 while a Stanford press release dated 19 March 2003 reported "over a million pages" indexed on the appliance.

Subsequent analysis of the query set revealed that it contained 83 queries for which the answer lay outside the `stanford.edu` domain (biased against the main Google index) and a further 33 within the `stanford.edu` domain but known to be excluded from the appliance index[8] (biased against the appliance). During the process of manual equivalence judging a number of pages were found to be inaccessible to us and these were also eliminated. Accordingly, analysis of the saved search results was repeated using only the remaining 1266 queries.

Table 1 shows the relative success achieved by site-restricted navigational queries on the main Google index and corresponding queries submitted to Stanford's Google appliance. In 907 out of 1266 queries, performance was identical; in 194 cases the appliance performed better; in 165 cases the difference was in favour of the whole-of-web service. A Wilcoxon matched-pairs signed ranks test shows no significant difference.

[7] `www.google.com/appliance/`

[8] `www.stanford.edu/services/websearch/Google/instructions4webcreators.html`, accessed 14 June 2003

Note that equivalence judging did alter the observed result. Without equivalence judging, the `google.com` performed slightly (4% on MRR1) but significantly better than the appliance. In other words, the Appliance more frequently returned equivalent URLs ahead of the ones listed in the site map than did the external service.

5 Experiments with seven Australian organisational web sites

In the absence of a complete Web crawl, we made use of three Australian crawls, covering the `.edu.au`, `.com.au` and `.gov.au` domains, totalling approximately six million pages. Crawls were carried out in late 2002 and early 2003.

For each of the chosen organisations we built two Panoptic indexes – one using local information only and the other using all the inlinks (and anchor text) originating from other domains within the aggregated crawls. In each run Panoptic was operated in a mode which ranking depends upon document content, document title, URL words, anchor text and inlink count (with offsite links weighted more heavily than onsite ones.)

Table 2 records details and results for the experiments involving Australian organisations and using the Panoptic search engine. In general, the extra evidence makes little difference. In total across the seven organisations, performance on 31 queries was improved and performance on 43 was made worse, out of a total of 1694.

Note that in most cases the crawls of the organisations were not run to completion and that with the exception of Curtin University, checks were not made to ensure that the correct answers were within the crawl.

NineMSN is a special case because there was no single site map and it was necessary to derive queries from a variety of link pages within the site. These pages were not removed before indexing. Initially 399 queries were collected but many of them referenced ephemeral content such as news items and became useless when the crawl corresponding to the laboriously generated query set was inadvertently overwritten. Results are reported only for the 98 queries whose answers were actually within the new crawl.

6 Discussion and follow-up

Our experiments show no consistent benefit from the addition of link evidence. Any benefit in individual cases is very small. These findings are perhaps surprising and it is worth trying to explain them.

We hypothesise that, in most cases, there is sufficient link evidence within an enterprise to identify the most important pages on a topic. In these cases, adding more evidence doesn't help. Furthermore, external links are heavily biased toward the main site entry page and a few key individual pages. They may offer little or no assistance in identifying important pages deep within a site. Finally, there is the possibility that external links may be slow to track changes within the site and may consequently point to outdated answers. The homepage of the main library at Curtin University provided an example at the time of our study[9] but we don't know how often this phenomenon occurs.

Table 3: URL-type breakdown for Australia Post. All figures are percentages of relevant population. Please note that some URLs classified as "Other non-param." may well be directory default pages – It is reasonably common for Web authors to omit trailing slash off links to directories.

		Site	Link targets	
URL type	Crawl	map	(Intern.)	(Extern.)
Vignette-style	33.9	97.9	11.3	3.4
Directory default	1.3	2.1	14.8	77.3
Other non-param.	20.4	0.0	36.2	7.9
Parameterised	44.3	0.0	37.8	11.3

For the organisations which we crawled, the number of external links to pages within the organisation was dwarfed by the number of internal links. CSIRO's external link count of 32,317 was less than 3% of the internal total of 1,209,028.

We have no inside knowledge of the algorithms used by the `google.com` service or the Google Appliance. It remains a valid interpretation of the Stanford results that the global service is disadvantaged by virtue of (a) inferior algorithm, (b) incomplete coverage of `stanford.edu` or (c) less frequent crawling, but is compensating by benefiting from the external link evidence.

The Australian experiments are not subject to these problems because both the search algorithm and the set of pages indexed are held constant. However, our supplies of external link information are incomplete for these experiments as we didn't have access to a complete crawl of the Web. Nonetheless, the sphere of influence of the organisations in question is Australia-oriented and it seems unlikely that a much larger crawl would substantially alter the pattern of incoming links.

The low confidence level criterion (95%) used in statistical testing, and the fact that eight tests were conducted, leads to a reasonably high probability of a spurious significant result. Correcting for this, by increasing the confidence level required, would make it even harder to reject the null hypothesis.

A study of the pattern of the known external links suggests that the availability of more external link information is unlikely to change the basic conclusion.

Web authors who create a link to the web site of another organisation tend to link either to that organisation's home page or to one of a few pages within the site which the author is likely to want to revisit or which are likely to be useful to other people. Australia Post illustrates this well. Of the external links to Australia Post, 45% are directed at the homepage of the organisation and 24% at the lookup service for postal codes.

Looking at the Australia Post distribution of URL types in Table 3 suggests that there are big differences between a) what is published, b) what is highlighted in the site map, c) what is linked to by internal pages and d) what is linked to from outside.

[9] The Curtin University library homepage moved but even though the bulk of internal links have been updated, a significant majority of external links still pointed to the old page.

Table 2: Results for Australian enterprise sites. The answer set for each query was expanded to include URLs which were detected by the crawler/duplicate eliminator to be duplicates of the correct answer or to redirect to a correct answer. In the cases marked `augmented`, the index included all available external link information. The site map page from which the query/answer set was derived was excluded in all cases except NineMSN (see text). "Number of changes", refers to the number of queries on which performance was improved ("+") or harmed ("-") by the addition of external evidence. Gain is shown as the percentage change in mean reciprocal rank of the first correct answer (MRR1) relatieve to the unaugmented run. An asterisk shows that the difference is significant on the Wilcoxon test ($p < .05$). None achieved significance at ($p < .01$).

	Additional Links	No. of queries	P@1	P@5	P@10	MRR1	No. of changes	Gain
Australia Post local	-	94	.4468	.7234	.7979	.5578		
Australia Post augmented	4175	94	.4362	.7021	.7979	.551	+0,-4	-1%
Comm. Bank	-	145	.5103	.669	.7241	.5794		
Comm. Bank augmented	1770	145	.5172	.6759	.7172	.5833	+3,-3	+.7%
Curtin U.	-	332	.5934	.7831	.8102	.674		
Curtin U. augmented	3478	332	.6175	.7892	.8163	.6895	+13,-6	+2%*
CSIRO	-	130	.3462	.5	.5462	.4179		
CSIRO augmented	32,317	130	.3538	.4923	.5462	.4217	+5,-7	+1%
DEST	-	62	.5000	.7903	.8710	.6234		
DEST augmented	14,902	62	.5161	.7903	.8710	.6348	+7,-11	+2%
NineMSN	-	98	.2959	.5204	.5816	.3848		
NineMSN augmented	7866	98	.2959	.5204	.5816	.3848	+0,-0	0%
UniMelb	-	532	.3327	.5639	.6335	.4308		
UniMelb augmented	15,607	532	.3308	.5639	.6335	.4290	+3,-16	-0.5%*

- Over 77% of external links reference the default page for a directory whereas only 2% of the site map entries do so.
- Nearly all the site map entries are to the Vignette part of the site but Vignette pages only constitute a third of the published content. By contrast, only 11.3% of internal and 3.4% of external links reference Vignette-format URLs (e.g. `www.auspost.com.au/BCP/0,1080,CH3345~MO19,00.html`).
- 44.3% of the site consists of parameterised ASP URLs (e.g. `www.auspost.com.au/philatelic/stamps/stampshop_2.asp?pid=679251689&product_type=8&category_id=291`), but none are linked to from the site map. This proportion is roughly reflected in the proportion of internal links, but external linkers are three times less likely to reference parameterised URLs.

7 Conclusions and Future Work

The novel evaluation methodology presented here is offered as a useful and low-cost tool for enterprise search evaluation. It is only applicable to navigational search and therefore can only be part of a toolkit, however in this particular study a focus on navigational search was desirable because that is where link evidence is most likely to pay off, and where the benefit of external evidence is likely to be greatest.

As noted in Section 3.2 the evaluation methodology is biased toward the publisher's view of a web site rather than that of visitors. However, alternative methods of query/answer collection are subject to problems which seem even more serious.

The enterprise webs we studied seemed to contain sufficient internal link evidence to allow good performance on navigational queries which make sense within the organisation. Any improvement to be gained on this task from the use of external links is very small. It is not considered possible to reject the null hypothesis.

In contrast, external link evidence may be important within a broader search context in ensuring that the organisation's home page is returned when the name of the organisation is used as a query. Links to the homepage from within the organisation may just use 'home' as anchor text.

Having shown that the performance of the Google appliance and of a site-restricted form of `www.google.com` are effectively indistinguishable on a navigational search task, it is reasonable to ask why an organisation would operate its own local search service. There are many possible reasons, including: greater coverage (several hundred thousand more pages indexed by the appliance at Stanford), greater control, greater freshness and ability to index internal, secure and non-web content.

Web authors external to an enterprise tend to link to the organisation's homepage and to useful services or information to which they or others are likely to return. Anchor text on such links reflects an external view of the purpose and value of the target resources. External authors favour directory default pages and short URLs and are much less likely to link to parameterised URLs.

Confirmation of the Australian results using a large global crawl awaits the availability of such a crawl.

References

E. Amitay and C. Paris. Automatically summarizing web sites - is there a way around it? In *ACM 9th International Conference on Information and Knowledge Management (CIKM 2000)*, Washington, DC, 2000.

Krishna Bharat, Bay-Wei Chang, Monika Henzinger, and

Matthias Ruhl. Who links to whom: Mining linkage between web sites. In *Proceedings of IEEE ICDM-01*, pages 51–58, 2001. `http://theory.lcs.mit.edu/~ruhl/papers/2001-icdm.html`.

Krishna Bharat and Monika R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of SIGIR'98*, pages 104–111, Melbourne, Australia, August 1998.

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of WWW7*, pages 107–117, 1998. `www7.scu.edu.au/programme/fullpapers/1921/com1921.htm`.

Andrei Broder. A taxonomy of web search. *ACM SIGIR Forum*, 36(2), 2002. `http://www.acm.org/sigir/forum/F2002/broder.pdf`.

Pável Calado, Berthier A. Ribeiro-Neto, Edleno Silva de Moura Nivio Ziviani and, and Ilmério Silva. Local versus global link information in the web. *ACM Transactions on Information Systems (TOIS)*, 21(1):42–63, 2003.

J. Carriere and R. Kazman. Webquery: Searching and visualizing the web through c onnectivity. In *Proceedings of WWW6*, 1997. `http://www.cgl.uwaterloo.ca/Projects/Vanish/webquery-1.html`.

Soumen Chakrabarti, Byron Dom, Prabhakar Raghavan, Sridhar Rajagopalan, David Gibson, and Jon Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of WWW7*, pages 65–74, Brisbane, 1998. `www7.scu.edu.au/programme/fullpapers/1898/com1898.html`.

Nick Craswell and David Hawking. Overview of TREC-2002 Web Track. In *Proceedings of TREC-2002*, Gaithersburg MD, November 2002. NIST special publication 500-251, `trec.nist.gov`.

Nick Craswell, David Hawking, and Stephen Robertson. Effective site finding using link anchor information. In *Proceedings of ACM SIGIR 2001*, pages 250–257, New Orleans, 2001. `www.ted.cmis.csiro.au/nickc/pubs/sigir01.pdf`.

B. Davison. Topical locality in the web. In *Proceedings of ACM SIGIR'2000*, pages 272–279, Athens, Greece, 2000. `http://www.cs.rutgers.edu/~davison/pubs/2000/sigir/`.

David Hawking and Nick Craswell. Overview of TREC-2001 Web Track. In *Proceedings of TREC-2001*, Gaithersburg MD, November 2001. NIST special publication 500-250, `trec.nist.gov`.

Jon Kleinberg. Authoritative sources in a hyperlinked environment. Technical Report RJ 10076, IBM, May 1997.

Henry E. Klugh. *Statistics: The Essentials for Research*. John Wiley, New York, 1970.

Lynx. Lynx browser home page. `lynx.browser.org`.

O. McBryan. GENVL and WWWW: Tools for taming the web. In *Proceedings of WWW1*, 1994.

Amit Singhal and Marcin Kaszkiel. A case study in web search using TREC algorithms. In *Proceedings of WWW10*, pages 708–716, Hong Kong, 2001. `www.www10.org/cdrom/papers/pdf/p317.pdf`.

Trystan Upstill, Nick Craswell, and David Hawking. Query-independent evidence in home page finding. *ACM Transactions on Information Systems (TOIS)*, 21(3):286–313, 2003.