Very Large Scale Retrieval and Web Search
(Preprint version)

David Hawking and Nick Craswell
CSIRO, GPO Box 664
Canberra ACT, Australia 2601
David.Hawking,Nick.Craswell@csiro.au

25 May 2004

Introduction

Together, the TREC Very Large Collection (VLC) Track and its successor the Web Track have run for seven years, after an initial VLC pre-track. During that time five new test collections have been created, five different types of retrieval task have been studied, a large number of important issues have been addressed, and new methods have been tried, not only for retrieval, but also for test collection construction.

Since the Web Track was a natural evolutionary step from the VLC Track, from here on we will refer to them as a single VLC/Web track.

The corpora created in support of the track have been distributed to more than 120 organisations world wide; they are clearly being used for evaluation and research purposes well beyond the confines of TREC. Not only that but the Web Track model has been adopted for similar Japanese language evaluations within the context of NTCIR (NII-NACSIS Test Collection for IR Systems, research.nii.ac.jp/ntcir/index-en.html).

Each edition of the VLC/Web Track (except the 1996 VLC Pre-track) has already been described in a Track Overview paper in the appropriate TREC Proceedings. [29, 26, 30, 20, 22, 23, 16] This chapter:

- Provides a layperson's guide to the track.
- Briefly summarises the history of the track and consolidates key information.
- Documents the VLC/Web resources which are available for ongoing experimentation and how to obtain them.
- Discusses the contributions of the track to achieving stated TREC goals.
- Lists the questions which have been addressed by the track over the years and outlines the current state of knowledge with respect to them.
- Discusses the impact which the track has had outside TREC.
- Reflects upon what has been achieved by the track and what has not.
- Recognizes the limitations on what can possibly be achieved by the track.
- Indicates directions for future work in the area.
- Acknowledges contributions.

Table 1: Summary of VLC and Web Track evaluations 1996 - 2003.

Year	No.	Track/Task	Coll.	Topics	No. Partic.
1996	TREC-5	Pre-VLC	CDs 1-4	251-300 (From Ad Hoc)	4
1997	TREC-6	VLC	VLC	301-350 (From Ad Hoc)	7
1998	TREC-7	VLC	VLC2	351-400 (From Ad Hoc)	7
1999	TREC-8	Large Web	VLC2	50/10000 (From SE NLQ logs)	8
		Small Web	WT2g	401-450 (Joint w. Ad Hoc)	17
2000	TREC-9	Large Web (Online Srvcs)	VLC2	50/10000 (From SE NLQ logs)	5
		Main Web	WT10g	451-500 (Rev. Eng. SE)	19
2001	TREC-2001	Web Topic Relevance	WT10g	501-550 (Rev. Eng. SE)	29
		Homepage Finding	WT10g	EP1-145 (Random target selection)	16
2002	TREC-2002	Topic Distillation	.GOV	551-600 (NIST engineered)	17
		Named Page Finding	.GOV	NP1-150 (NIST engineered)	18
2003	TREC-2003	Topic Distillation	.GOV	TD1-TD50 (NIST engineered)	23
		Mixed Named/Homepage	.GOV	NP151-450 (NIST engineered)	19
		Interactive (Topic Dist.)	.GOV		2

Layperson's guide to the track.

The initial VLC Track evaluations were very similar to those in the Ad Hoc task described in Chapter [REFER TO CHAPTER 4] and used the same NIST-constructed topics. As time went by, the track focused more on web search and diverged further from the Ad Hoc formula.

Web search is different from the retrieval modelled by TREC Ad Hoc because of the size of the data sets (up to five billion documents or more), the type of document, the presence and nature of interlinking between documents, the volume of queries submitted (around five hundred million queries per day to Web search engines), the length of typical queries (a little over two words on average) and the types of search activity undertaken.

Web terminology

Some web-specific terms deserve explanation. We use the term *Web* to refer to the World Wide Web and *web* to refer to any hyperlinked collection of documents served by web protocols, particularly the HyperText Transfer Protocol (HTTP) [33]. The Web is an example of a web and a web may or may not be a subset of the Web!

Generally, web documents are encoded in HyperText Markup Language (HTML - See`www.w3.org/MarkUp/Overview.html`.) though they may also link to images and documents in other formats. Conventionally, web documents are addressible via a unique address in the form of a Uniform Resource Identifier (URI), more commonly known as a Uniform Resource Locator (URL). [43] An HTML document accessible by this means is usually called a *web page* or just a *page*. Within an HTML document certain groups of words serve as the *anchors* of outgoing links. These words are normally highlighted when the document is displayed in a browser.

The collection of text from the anchors of all links targeting a particular page is called its *referring anchor text*.

As well as the structure imposed by the hyperlink graph, a web possesses structure because of the

relationship between URLs. This gives rise to the concept of *web site* or just *site*. For example all the pages whose URL start with `trec.nist.gov` may be considered to be on the "TREC site" regardless of links. Those pages which start with `trec.nist.gov/pubs/` may be considered to be on the "TREC publications" site. Sites usually have an *entry page* which can also be called a *home page*. The URL of a home page is likely to end with a slash `trec.nist.gov/pubs/` or in one of a small number of names such as `index.html` or `default.htm`.

The Web is highly dynamic and, due to automatic page generators, infinite. Pages which are not linked to by any other page are discoverable only by knowing or guessing the URL. Such pages may or may not be considered part of the Web (depending upon definition). It is not possible to take a complete snapshot of the Web because the only way of identifying "all" the pages is to use a *crawler* (also known as a robot or spider). Crawling can never discover all pages and takes weeks, during which time the structure and content of the Web will have changed significantly.

A crawler starts with a list of to-visit URLs (perhaps just `yahoo.com`). It operates by taking a URL off the list, adding it to a list of URLs which have been visited and fetching the page at that URL. Once a page is fetched, its links are extracted and any previously unseen URLs are added to the to-visit list. This process continues until the to-visit list is empty. The collection of pages fetched by a crawler is referred to as *a crawl*. Corpora used in the Web Track are crawls or selections from them.

Sometimes crawlers are configured to fetch only URLs within a specified Internet domain (or set of domains) in order to provide enterprise-level search. For example, an outward facing local search engine for Sony might crawl only the publicly accessible pages within `sony.com`. Behind the Sony firewall may be another *enterprise web*, inaccessible to the outside world but searchable by the company's *enterprise search engine*. In general, an enterprise search engine must search not only internal web documents but email, database records, contents of document management systems and files on shared hard drives.

Other crawlers may visit pages determined to be more likely to be relevant to a particular topic. This is called *focused crawling* and may find application in a *subject portal*, providing a single search interface to Web resources on a topic such as chemistry or mental health.

Web *search engines* include a crawler as well as a conventional text retrieval system. *Metasearchers* are brokers which broadcast queries to a set of primary search engines and merge the results.

Typical tasks in the track

A typical Web Track experiment proceeds as follows:

1. **Documents and query topics:** Participants are provided with a set of Web documents and a set of fifty or more query topics. Each query topic supplies a user's query ("Qantas") and either states or implies an underlying need ("find me the Qantas home page", or "find me all pages about Qantas", or "find me a short list of the most important Qantas pages").
2. **Submitting runs:** Participants run the queries over the documents and submit the top n results to NIST. The value of n depends upon the task.
3. **Judging (Informational/Transactional):** NIST creates a document pool for each query, based on the results of all participants. Assessors are employed to judge which pooled documents would satisfy the user's underlying need. In the case of known (or suspected) item search, the correct answer is determined in advance and the only requirement for human judging may be to identify effective duplicates of it.
4. **Judging (Navigational):** Assessors are employed to locate target items such as homepages. Judging is then only required to identify duplicates of the target items within the submitted runs.

5. **Evaluation:** The set of documents judged to satisfy each query is made available. Based on these judgments, effectiveness can be measured in a number of ways. Typical Web Track measures are:

    - Precision at n (P@n): Proportion of the top n documents which are satisfactory. P@10=.4 means that 4 of the top 10 documents were satisfactory. A run of 50 queries would be measured according to mean P@n.
    - Success at n (S@n): The proportion of queries for which a correct answer was within the top n. S@1=.5 means that for half the queries, the correct answer was at rank 1.
    - Mean reciprocal rank of first correct answer (MRR1).
    - Mean average precision (MAP).

If method A is significantly better than method B over a large enough number of queries (and, even better, over multiple TRECs) then we believe that method A is superior for this task. For example, home page finding effectiveness has been improved in a number of experiments using anchor text propagation [15] or URL type classification [32] so we believe these are useful techniques for home page finding.

It is understood by VLC/Web Track organisers and participants that real web search is a very complex and dynamic human activity usually undertaken as part of some broader task. However, the VLC/Web Track has mostly focused on trying to maximize the value of each individual query-response transaction between a searcher and the search engine. In 2003, this focus was broadened with the incorporation of former Interactive Track activities as a sub-track.

VLC/Web Track evaluations are conducted in order to learn things which will help make more useful search systems for use in the real world. Each of the ingredients of a Web Track experiment should therefore be representative of a real-world application. The documents should represent a real document set. (For example, a crawl of `.gov` as might be used by a US Government search engine.) The query topics and assessors should be representative of real user needs and preferences. Evaluation measures should accurately reflect real user requirements and behaviour. For example, P@1000 isn't a very realistic measure in Web search because few Web searchers look at more than the first five or ten results.

Potted History of the Track

Table 1 provides a summary of dates, collections, tasks and participation. The following brief history provides some explanation and context for the events.

In November 1995, David Hawking and the late Paul Thistlewaite proposed the creation of a Very Large Collection Track, in order to ensure that TREC kept pace with burgeoning text collections, particularly the Web. The track was intended to provide a focus within TREC for the study of scalability, efficiency and the applications of parallelism.

Two new corpora were created to serve these objectives. The first was the 20gB VLC, released in 1997. It included large quantities of newspaper and government data, many gigabytes of USENET news and a small amount of Web data. The second was the 100gB VLC2, a truncated Internet Archive [31] Web crawl from February 1997. It was released in 1998 and represented a 50-fold increase in data size over TREC Ad Hoc. Both VLC and VLC2 contained too many documents to justify the assumption made in TREC pooling of *sufficiently complete* relevance judgments. Accordingly, the evaluation focus with these collections was necessarily, and appropriately, given the nature of typical Web search, on early precision.

In 1999, attention shifted away from efficiency and scalability and toward conducting evaluations which simulated more of the features of Web search. The Large Web Task in 1999 required the processing of 10,000 queries extracted from real Web search logs, of which 50 were selected post hoc for judgment.

The 1999 Small Web Task addressed the question of whether hyperlink information could be used to improve ad hoc retrieval effectiveness. The Small Web Task used a subset of VLC2 documents small enough to allow easy participation and to enable sufficiently complete relevance judgments.

At the Infonortics Search Engines Meeting in April 2000 (`www.infonortics.com/searchengines/sh00/boston2000pro.html`), Chris Buckley and David Hawking argued the case for the application of TREC evaluation methodology to the development of Web search engines. A spirited debate between the TRECcers and a panel comprising Larry Page (Google), Eric Brewer (Inktomi), Marc Krellenstein (Northern Light), Andrei Broder (Alta Vista) and Jan Pedersen (recently of InfoSeek) was (in the opinion of a fully engaged participant) quite valuable to both sides. The search engine representatives argued that the bulk of Web searches did not correspond to the *seeking a range of relevant information* task model assumed in TREC ad hoc (and in the Web Track to that time). They also argued that "relevance" judgments should record multiple levels and use judging criteria appropriate to the task. The Web Track responded quickly to these suggestions.

In 2000, queries were selected for judgment in the Large Web Task on the basis of whether they seemed to be attempts to locate an online service, such as downloading MP3 files or sending flowers. Documents were judged to be *useful* if and only if they provided direct access to the desired service. (Relevance was not sufficient.)

Also in 2000, three-level relevance judgments were used for the first time in the Main Web Task. Contrary to her own expectations, Voorhees [44] showed that the ranking of runs did depend upon whether they were evaluated using all relevant documents or highly relevant documents only. The topics used in the Main Web Task were reverse engineered by NIST assessors. They chose queries of interest from a Web log and used those as the titles of topics. They then decided upon an interpretation of the need behind the query and filled in description and narrative fields accordingly. In 2000, some of the titles deliberately included misspellings.

The 2000 Main Web Task used a new, carefully engineered selection from VLC2, called WT10g [5]. It was selected to ensure a high proportion of inter-server links. Despite this, participants did not report major gains in ad hoc effectiveness through use of links.

In 2001, use of the VLC2 corpus was suspended because it was felt that more could be learned from WT10g. The previous years Topic Relevance Task was repeated with very similar methodology, but spelling errors were not included this time. In addition, a new type of search (homepage finding) was introduced. The judging criterion in this case was whether a document was the home page (site entry page) of the entity named in the query. In this type of search both URL and link structure were found to be highly beneficial.

In 2002, two further types of search have been introduced and evaluated in the context of the .GOV collection, Named Page finding and Topic Distillation. Unlike the WT2g and WT10g collections, which were artificial selections from a large whole-of-Web crawl, .GOV is a natural (albeit truncated) crawl of a limited but interesting Internet domain. .GOV is a 1.25 million page crawl of the .gov Internet domain collected in early 2002.

Note that homepages constitute the only correct answers in navigational searches but are also very valuable in informational searches – if someone is searching for information about NIST, they will be happy to see NIST's home page at the top of the list.

The Named Page Finding Task was a simple variation of the Homepage Finding Task. For each query, the desired result is a single important document but it was not in general the entry page of a website.

Both named page finding and homepage finding are related to known-item search in that in each case a single answer is sought. However there are important differences. The usual scenario underlying known-item search is that the searcher wishes to re-locate a document they have seen before. In home-

page finding, it is frequently the case that the searcher has never visited the target document before and merely suspects that it exists – *suspected item search*. Not only that, but the homepage may not contain any text which matches the query or indeed any text at all. It is also possible that a named page has not been visited previously (e.g. bus timetable for service 359) and may sometimes be graphical rather than textual (e.g. `A-Z map of London SE1`).

The Topic Distillation Task introduced in 2002 was related to earlier Topic Relevance tasks, but intended to identify key resources on a topic. The task proved difficult to explain precisely to both participants and assessors. In essence a perfect topic distillation system would, when given a broad topic, make a short list of key resources which would closely match the list a human might create as a bookmark file or to give to students of the topic. Definition of the ideal list is, like relevance, in the mind of the human judge. Unfortunately, documents identified by the 2002 judges as being key resources did not accord very well with what participants were expecting.

Accordingly, in 2003, the Topic Distillation task was simplified. Key resources were constrained to be web sites, represented by their entry pages. Thus the topic distillation task became a type of homepage finding, with the important distinction that the query is not generally the name of the entity represented by the web site, and there will usually be several good homepages for a topic.

In 2003, a combined Home Page/Named Page task was set with a total of 300 topics. Also in 2003, the Interactive Track administratively became a sub-track of the Web Track and studied human performance on topic distillation.

Relation to TREC Goals

The VLC/Web Track has contributed significantly in achieving TREC's four goals. [36]

Goal 1: To encourage research in information retrieval based on large test collections

We interpret *large test collections* as meaning, "large enough to be confident that results obtained will apply in the majority of current retrieval applications of that type." We consider it risky to extrapolate results from a test collection of a particular collection to real collections more than an order of magnitude larger. The definition of what qualifies as a *large test collection* has needed frequent review because the lifetime of TREC has coincided with phenomenal growth, not only in the scale of real search applications but also in the volume of searches conducted.

In 1992 the scale of the TREC-1 ad hoc collection reflected that of CD-ROMs, electronic newspaper archives, collections of legislation, parliamentary transcripts and the electronic text holdings of government agencies and certain enterprises. Library catalogs were of a similar scale and so were information sources on the Internet such as FTP and Gopher servers and USENET news collections. The total amount of electronic text spread across these types of collection was of course enormous but, in 1992, there was no effective means to combine them.

Even at the time of TREC-1, powerful forces were at work which would inexorably increase the scale of text collections over which people wanted to search. Even a steady rate of accumulation of new material resulted in high proportional growth because many organisations had only recently started to store their text electronically. In addition, the number of organisations and individuals producing and publishing electronic text was increasing and means were being developed by which collections could be aggregated for the purpose of unified search.

The Wide Area Information Service WAIS was released by Thinking Machines Corporation in 1991 [46] and by February 1993, there were approximately 400 WAIS servers. (`www.upenn.edu/computing/printout/archive/v09/4/navigation.html`). The University of Nevada's

VERONICA tool for searching multiple Gopher Sites was first released in November 1992 and that same year saw the first implementations of the Z39.50 protocol. [35] However, by far the most significant development was the advent of the World Wide Web.

In early 1992 [8] the first Web browser was released and late in that year there were 26 generally accessible Web servers. [6] Within the lifetime of TREC that number has grown to an estimated 150 million! [46] Furthermore, the world has come to expect that all of those sites will be searchable via a single interface.

The Lycos search engine was launched in May 1994 by Michael Mauldin of CMU (`www.clubi.ie/webserch/engines/lycos/index.htm`) and AltaVista followed in December 1995. The subsequent growth in the amount of Web data indexed by commercial search engines was so rapid that TREC could not keep up.

At the time it was first distributed (1998) the 18.5 million page VLC2 collection was comparable to the coverage of Lycos, though much smaller than that of Alta Vista. Five years later, in 2003, major search engines were indexing 200 times as many documents as are in VLC2!

Given that the average size of Web pages in 2004 is of the order of 20kB, the data size of the collections indexed by Google and Yahoo! search engines is of the order of 60 - 100 terabytes, not including images and other binary data. Distributing that quantity of data as a test collection would be logistically infeasible. Not only that, but few if any TREC participants would be able to deploy the hardware resources and advanced engineering capability necessary to deal with it.

While TREC is constrained to operate with test collections a couple of orders of magnitude smaller than the Web, Web search engine companies are unlikely to look to it for advice on scalability or efficiency techniques. Furthermore, results obtained using WT10g or .GOV scale data sets and link graphs should be extrapolated with great caution to the Web as a whole. Despite this, the VLC/Web Track does offer a collaborative environment in which methods and hypotheses can be proposed and tested.

Retrieval within the webs operated by organisations is both commercially important and scientifically interesting. The scale of VLC/Web evaluations is well matched to this type of problem. The amount of text data held by large organisations is growing but it remains true that few individual organisations hold more than 18.5 million documents.

Diminishing difficulties posed by large test collections.

In 1992 the 2 gB data size of the TREC-1 collection set a difficult indexing hurdle which some participants were unable to surmount. [19] However, hardware and indexing advances over the next few years dramatically reduced the challenge.

At TREC-7 (November 1998), David Hawking showed that it was possible to index the Ad Hoc collection and process 50 queries within a one-hour conference slot, using only a mid-range laptop computer (266MHz Pentium II, 128MB RAM and 6gB internal disk). The following November he demonstrated that with the addition of a second 10gB disk, the same laptop could process queries at a reasonable rate over pre-built indexes for the full VLC2 collection.

In TREC-2002, 23 groups succeeded in indexing the 18 gB .GOV collection and submitting runs.

Goal 2: To increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas

The Web Track attracted more than ten commercial participants in 2001. However, endeavours to persuade [27] and later to provoke [24] Web search engine companies to join in have so far been unsuccessful.

Despite this, employees of commercial search engine companies such as Jack Xu (Excite), Andrei Broder, Monika Henzinger and Michael Moricz, Peter Anick and Bob Travis (Alta Vista), Knut Magne Risvik and Per Gunnar Auran (Fast/AlltheWeb), Edwin Cooper (the Electric Monk), Krishna Bharat, Amit Singhal, Larry Page and Ron Dolin (Google), Raman Chandrasekhar and Bill Bliss (MSN Search), and Andy MacFarlane (Omsee) have contributed to the planning of track activities and interpretation of results. (Affiliations shown are those applicable at the time.)

An invited talk by Andrei Broder (then Chief Scientist at Alta Vista) during the Web plenary session at TREC-9 (2000) [7] in which he proposed the classification of search types into Informational, Navigational and Transactional and expanded on his Search Engines 2000 coining of the term "Adversarial Information Retrieval", was very well received by the TREC audience.

A number of attempts have been made to highlight the relevance of TREC results and methodology to the search engine companies by comparing the performance of TREC systems with that of public search engines (without their cooperation). These are summarised in the section (below) entitled "How do TREC systems compare to Web search engines?".

It is now clear that the major search engine companies take scientific evaluation of their search quality very seriously, even if they do not publish results of their evaluations.

Goal 3: To speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems

It is difficult to estimate the rate of technology transfer between TREC and the Web (and enterprise) search industry because commercial companies are free to adopt directions, algorithms and methodologies published in TREC and under no obligation to disclose to the extent to which they do.

As mentioned above, there are significant limits on how much the VLC/Web Track can influence the engineering of whole-of-Web search engines. It is quite clear that major search engine companies are a long way ahead of academic research laboratories in these areas. They have accumulated substantial experience with crawling and indexing quantities of data orders of magnitude beyond the scope of TREC.

Furthermore *crawl quality* and *spam rejection* are essential to high quality Web search and neither are amenable to study within a static test collection. The quality of a crawl refers to its recency and coverage and the extent to which low-value, and duplicate or near-duplicate pages are excluded. Spam refers to artificial Web pages and structures designed to inappropriately promote certain pages in search engine result lists. To be successful, spam technology must and does adapt rapidly to changes in search engine algorithms. Any counter-measure developed to defeat one type of spam would in reality be immediately subject to a counter-counter-measure, but this cannot happen in a static test collection.

As previously noted, the web search industry is much broader than whole-of-Web search and the VLC/Web evaluations have direct bearing on enterprise and portal search. It is known that technologies based on VLC/Web Track results do make their way into products in these areas. (See Page 20.)

Goal 4: To increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems

This is an area in which major progress has been made by, and in conjunction, with the VLC/Web Track. As noted in the introduction five test corpora (see Table 2) have been created and evaluation methodologies for five different types of search (topic relevance, online service finding, homepage finding, named page finding and topic distillation) have been developed.

Judgments involving WT2g, WT10g and .GOV are for most purposes re-usable. Although past judgments involving VLC and VLC2 are incomplete and therefore not re-usable, the cost of rejudging sufficient documents to make P@20 (precision at 20 documents retrieved) comparisons is low, even if assessors are paid.

Homepage finding and named page finding judgments are both cheap to create and reusable as, potentially, are those for topic distillation. Manual judging may be required to identify answers which are duplicates of the listed one or which automatically redirect to it, but judging whether two pages are identical is far simpler than assessing whether their content is relevant to a topic. In any case the judging process is largely capable of automation.

Issues addressed by the VLC/Web Track

Rather than repeating TREC-by-TREC material from the VLC/Web Track overviews and participant reports, we draw out the issues which were addressed and summarise the findings.

Scalability

The TREC-6 and TREC-7 VLC Tracks investigated the scalability of retrieval systems on various dimensions including query processing time, index size and index building time. Unsurprisingly, index size tended to grow linearly with collection size but the scalability of the time measures was very dependent upon the nature of the particular system and the hardware employed.

Engineering issues

Well-known principles apply in engineering IR systems for high performance:

1. use RAM rather than disk (design data structures to ensure high memory reference locality, use compact structures and compression to make best use of limited RAM resources),
2. use cache rather than RAM (use tight loops in frequently executed code),
3. use efficient algorithms,
4. consider potentially lossy optimisations, such as early termination,
5. if applicable, minimize communication between processors,

but, as with the scalability, the question of which engineering issues are most important is system and hardware dependent.

In the TREC-7 VLC Track, the ACSys VLC Medal was awarded to the University of Waterloo group for a) indexing the VLC2 in less than 10 hours (8.53), and b) processing queries in an average of under 2 seconds (0.882), while c) achieving median P@20 or better. They used a cluster of four PCs costing a total of around $US8,500.

In the TREC-8 Large Web Task, an attempt was made to explore the tradeoffs made by the participating systems across five key measures:

1. Speed of indexing;
2. Size of indexes;
3. Speed of query processing;
4. Query processing effectiveness; and

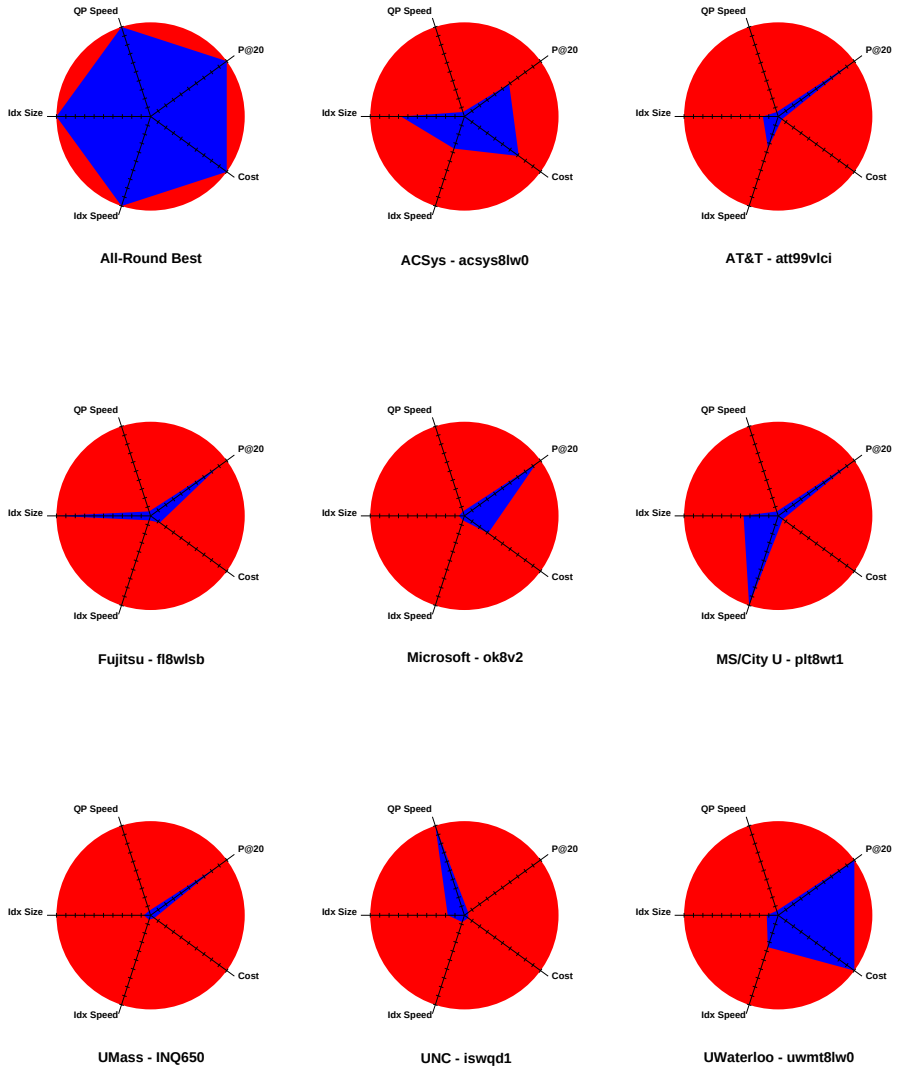

Figure 1: Composite results for illustrative runs submitted in the TREC-8 Large Web Task. Note that the UNC runs were submitted after the deadline and consequently included a very high percentage of unjudged documents. Accordingly, their precision result is very low. However, their query processing was two orders of magnitude faster than the next fastest, scaling other speed results into oblivion. The AT&T run was also unjudged due to a formatting problem. The All-Round Best is a hypothetical composition of the best-achieved result on each dimension. Finally, because ACSys co-ordinated the track, employed assessors and tabulated results, ACSys results should be regarded as unofficial.

5. Cost.

Kiviat diagrams were chosen to communicate the tradeoffs. Figure 1 shows 5-axis Kiviat diagrams summarising the performance on each of these dimensions of several TREC-8 runs. On each axis, best performance is represented by a point on the circumference. For effectiveness, best performance corresponds to maximum P@20 score whereas in each other case best performance corresponds to minimum score.

To illustrate the scaling process, the smallest index size was achieved by Fujitsu at 3.9 gigabytes. (They preprocessed the data to remove binary and non-English data.) This minimum was divided by the actual index size for each run to give a scaled score of 1 for Fujitsu and a score of 0.1 for a hypothetical index of 39 gigabytes. Scaled scores of less than 0.05 are shown as 0.05 to prevent the creation of spikes which are too narrow to see.

In TREC-8, the University of Waterloo team demonstrated sub-second query processing over VLC2 on a pair of cheap PCs which they brought along to the conference. In TREC-9, Fujitsu Laboratories set a new mark for the cheapest system used to run the Large Web task. With a $US1700 dual-Celeron system (648 MB RAM, and 3 x 40gB disks) they indexed the data in just over 12 hours (including decompression) and were able to process queries in an average of 0.31 sec.

Effect of collection size on effectiveness

Contrary to the expectations of many, e.g. Salton and McGill [38, p. 173], the observation of all participants in the TREC-7 VLC Track was that P@20 was considerably greater for retrieval over the full VLC collection compared with retrieval over a 10% sample of it. Various hypotheses were advanced as to why this might be so and these have been analysed in considerable detail by Hawking and Robertson [28] who found that observed behaviour could be well described by a signal detection (SD) model with due allowance for discreteness.

How does Web data differ from Ad Hoc data

In TREC-8, the question of whether the effectiveness of retrieval systems was dependent upon the type of data was studied in the Small Web task. The same topics were used in both the Ad Hoc Track and in Small Web and they had been developed with both collections in mind. The TREC-8 Web Track overview compares the mean average precision for each of ten matched pairs of runs across Small Web and Ad Hoc. For convenience the resulting scatter plot is reproduced in Figure 2. The Pearson R coefficient of correlation is 0.884, which is significant at the 0.05 level (two-tailed).

Are hyperlinks useful in topic relevance tasks?

Quite a few groups have attempted to demonstrate a benefit of using link information in Topic Relevance Tasks (i.e. the well-known TREC Ad Hoc task in which the goal is to find "find as many documents as possible providing information relevant to a topic").

In TREC-8, 20 out of 44 runs in the Small Web Task (WT2g) attempted to exploit link evidence. Methods employed included sibling pages, Kleinberg hub and authority (and variants), PageRank, spreading activation, probabilistic argumentation systems, indegree and outdegree. The differences between content-plus-link runs and the corresponding baseline were mostly very small and usually negative. The few large differences were all negative. Both the University of Neuchatel and Fujitsu Laboratories reported that they could find no correlation between relevance on the TREC-8 topics and link-based measures. ACSys found no benefit in the Large Web task from the use of PageRank scores.

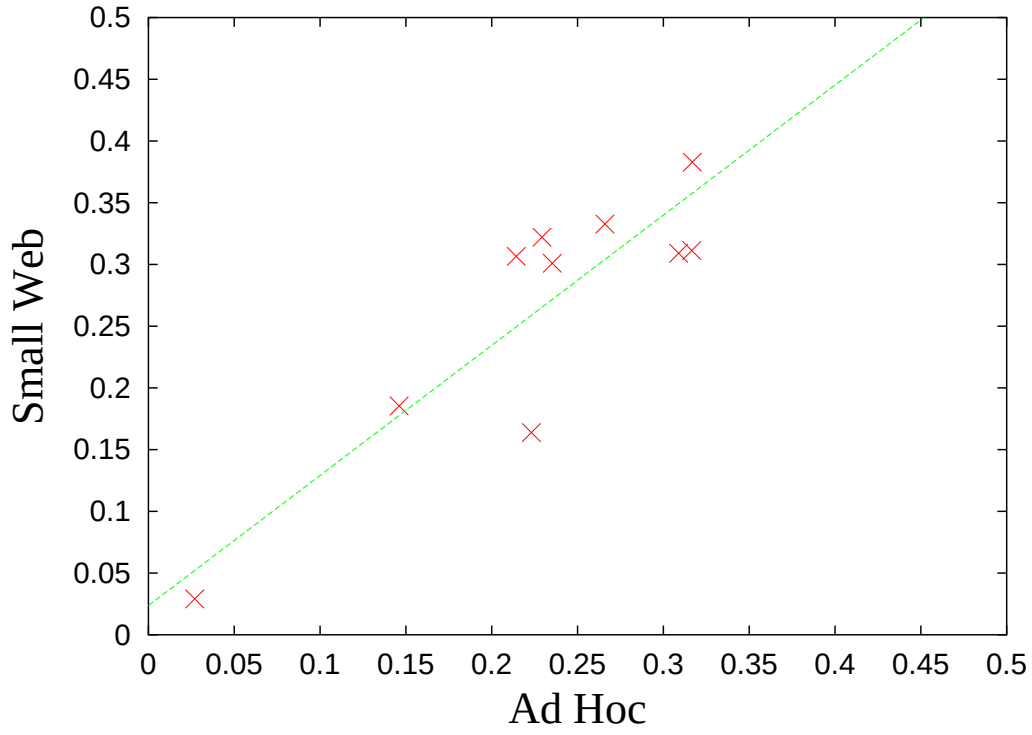

Figure 2: Average precision on the TREC-8 Small Web Task plotted against average precisions on the TREC-8 Ad Hoc task for pairs of runs believed to correspond closely. Also shown is the line of best (least-squares) fit.

A strong motivation in the replacement of WT2g with WT10g in TREC-9 was the low number of inter-server links within the former collection and the possibility that this was responsible for the negative results of the link trials. However, the proportion of TREC-9 Main Web Task submissions which attempted to exploit links dropped to 27 out of 105. Once again, despite the richness of the hyperlink graph in this artificial Web susbset, gains, if any, due to link methods (including anchor text) were inconsistent and at best very small. This was true even when highly relevant documents were valued very highly.

The exploration of link methods within the TREC-2001 Topic Relevance Task was less vigorously pursued. Once again it was shown to be possible to achieve top results using only document content.

To many, this is a very surprising outcome and one which challenges the Web Track's claim to be doing interesting work. Many find it beyond question that hyperlink methods work because a) Google uses hyperlink methods and b) Google produces good results.

The resolution of this apparent dilemna is straightforward:

> *Hyperlink and other web evidence is highly valuable for some types of search task, but not for others. Because binary judgements were employed and judges looked only at the text of the retrieved pages, the TREC-8 Small Web Task and the TREC-9 Main Web Task did not accurately model typical Web search.*

In the next section we cite evidence that, using TREC Ad Hoc evaluation methodology, TREC systems (without hyperlink evidence) actually outperform well-known Web search engines on whole-of-

Web search. But this is a relatively meaningless victory since, as was forcefully stated by the panel of Web search experts at the 2000 Infonortics Search Engines Meeting `www.infonortics.com/searchengines/sh00/boston2000pro.html`, the TREC Ad Hoc search task [REFER CHAPTER 4] is not at all typical of search on the Web.

In prototypical TREC Ad Hoc methodology, the task presupposes a desire to read text relevant to a fairly precisely defined topic, and documents are judged on their own text content alone as either relevant or not relevant.

By contrast, Web searchers typically prefer the entry page of a well-known topical site to an isolated piece of text, no matter how relevant. For example, the NASA home page would be considered a more valuable answer to the query 'space exploration' than newswire articles about Jupiter probes or NASA funding cuts. As a further example, there are estimated to be around 40 million webpages matching the query Microsoft, but web searchers expect that the entry page to the official Microsoft site will be ranked first.

People search on the Web for a variety of reasons in which a long unordered list of matching documents is not useful. They may wish to visit a site where they can browse or perform local searches, they may wish to find contact details or to answer a question, they may wish to buy something, or alternatively they may wish to access an online service.

A number of types of search task are listed on Page 14 modelled within the Web Track in which web evidence (such as anchor text, URL structure and hyperlink measures) brings dramatic benefits. The measures used in these tasks reflect the fact that certain pages, particularly site entry pages, are much more valuable than isolated matching pages.

Web measures are effective at ranking on likely value to the searcher, within the set of relevant documents. Note that in Web search literature the term *value* is not typically used. Instead, people have written about importance, popularity, authority etc.

How do TREC systems compare to Web Search Engines?

Two studies involving the present authors [26, 27, 25] showed a significant superiority of TREC systems over commercial Web search engines on a Topic Relevance task. In these comparisons retrieval systems in the TREC-7 VLC Track were considered to be search engines all sharing the same incomplete and out of date crawl. The same queries were fed to a number of public search engines and the results were pooled and judged by the same judges who evaluated the VLC submissions.

We also collected the data necessary to compare TREC systems with public search engines on the TREC-9 Online Service location task, but did not publish the comparison. That oversight is rectified here in Figure 3. The mean precision of the search engines (0.4932) is about 9% higher than the mean for the best runs from each group. However, the victory to the search engines is hardly decisive as the decision to evaluate online service finding for the TREC systems was made post hoc, after runs had been submitted.

By contrast, Singhal and Kaszkiel [41] compared a well qualified TREC system against public search engines on a homepage finding task and found that the latter were greatly superior.

The difference in outcome of these four studies, highlights to us the importance of modelling different types of search. At the time of the Singhal and Kaszkiel study, TREC systems were not oriented toward homepage finding and did not make use of link and URL evidence later found to be very beneficial on this task.

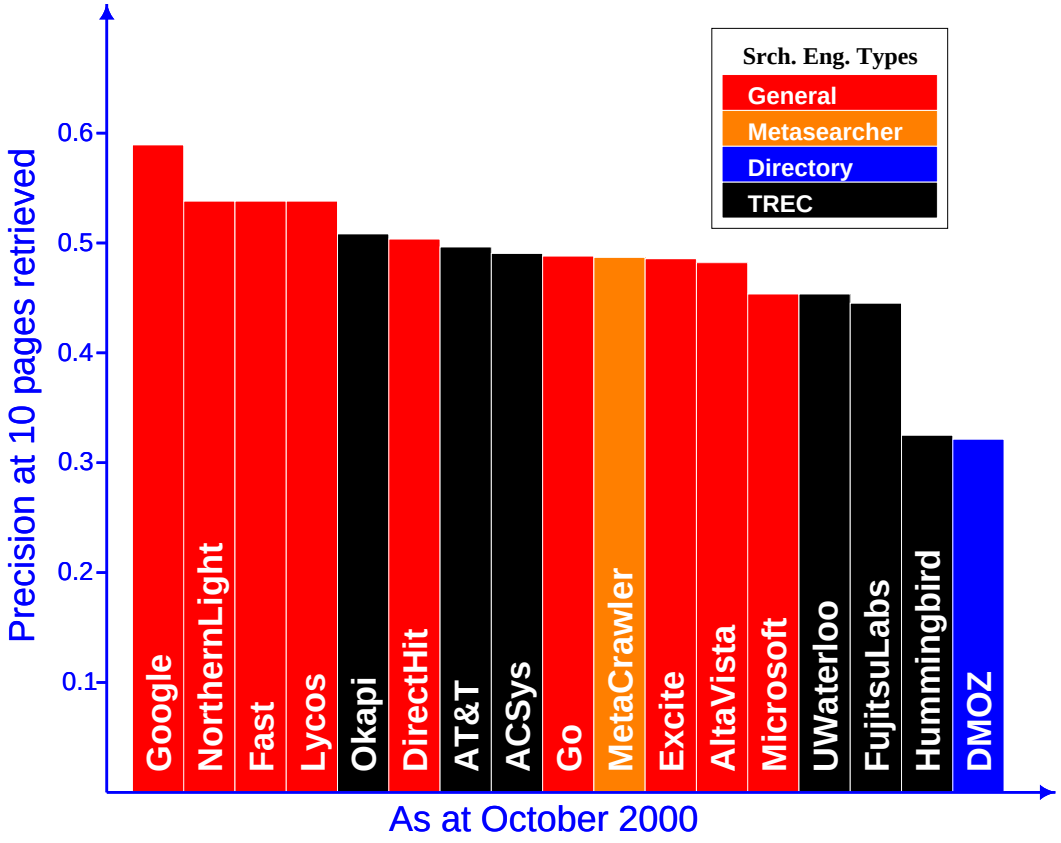

Figure 3: P@10 for 84 online service queries, comparing TREC Large Web systems with public search engines.

Different types of search

As noted above, web search takes many forms. The VLC/Web track has explored quite a few.

Online service finding

Online service finding is obviously a commercially important type of search which is not well understood. As noted earlier, the adoption of this task in TREC-9 Large Web was decided post-submission and there was no opportunity to tune systems or compare methods. Furthermore, it is not at all clear that VLC2 was a good testbed for this type of search.

Future work in this area may well involve a specially constructed corpus comprising commerce sites and appropriate reviews and referral agencies.

Homepage Finding

On the Home Page Finding task in TREC-2001, web-specific methods, specifically link anchor text and URL structure, came into their own. Referring anchor text was shown to be highly beneficial. Craswell et al [15] and Upstill et al [42] have confirmed this on several different collections. Attempts to use link graph measure such as inlink counts and PageRank were not as successful. Best results on the TREC

task were achieved by TNO/UTwente who calculated prior probabilities of different categories of URL (root, subroot, directory default and file). [45, 32] This was a previously unpublished method and, as far as we know, is not widely used by search engine companies.

Named Page Finding

Named Page Finding was proposed as a task to see if significant pages which were not site entry pages also tended to attract links or to be distinguished by the nature of their URLs. The two top performing runs in the TREC-2002 task used a fusion of different methods, including anchortext, but did not gain benefit from URL structure, URL length or inlink count. [47, 13].

Mixed Named Page / Homepage

In TREC-2003, 300 navigational queries were generated by NIST, comprising a 50/50 mixture of homepages and named pages. Overall, it seemed useful to consider different document representations or surrogates (such as referring anchor text) and to fuse results. Link structure gave mixed benefit. A number of attempts to switch retrieval methods based on automatic classification of query type showed promise but didn't surpass the best other methods.

Topic Distillation

Topic Distillation was proposed as a task to see if Web-specific features such as links and URL structure could be used to distinguish key resources from among the set of pages relevant to a topic. By key resources, we meant the type of resources which a human editor might list under a subject category in the DMOZ, Yahoo! or LookSmart online Web directories.

If a purely automatic system could produce a high-quality DMOZ-like directory in response to a subject category as query, not only would there be a major potential saving in human effort, but searchers would not be restricted to categories previously identified.

Unfortunately, in TREC-2002, the intended nature of Topic Distillation was not effectively communicated by the organisers to assessors and participants. As a result the topics chosen sometimes corresponded to no key sites within .GOV. Indeed, the lists of key resources were qualitatively different to DMOZ lists and tended to be isolated pages which the assessors felt were highly significant.

Topics were defined in standard TREC format and included a description, e.g.

```
<top>
<num> Number: TD26
<title>Nuclear power plants</title>
<desc>Description:
Operational and safety information associated with nuclear power
plants.
</top>
```

Sometimes, as in the example shown, the description tended to narrow the definition of the topic, encouraging the choice of single documents rather than sites.

Results for 2002 were inconclusive due to confusion about the task definition. Participants who developed algorithms biased in favour of retrieving sites rather than individual pages were unable to say whether their methods were effective.

In TREC-2003, the Topic Distillation task was redefined. This time systems were required to find entry pages to web sites devoted to specified broad topics. This definition no doubt excluded some key single-page resources but was well understood by participants and assessors and enabled participants to focus on web-specific attributes of the problem.

To illustrate, the official answers to the topic 'cotton industry' included homepages of the Cotton Pathology Research Unit, the FAS Cotton Group, the Western Cotton Research Laboratory and the USDA Cotton Program.

Readers are referred to [16] for details of experiments and conclusions. In general, referring anchor-text was found to be useful, and URL structure and link counts were also helpful.

A number of participants noted that the TREC-2003 topic distillation task was quite representative of real Web search.

Evaluation methodology for Web retrieval

Many of the methodological decisions taken in the Web Track are relatively uncontroversial but some questions have not been fully resolved. Issues in the evaluation of public search engines are addressed in [18, 14] and [25]. The key issue of using measures and judging instructions appropriate to the type of search, applies equally in test collection evaluations.

Evaluating the effectiveness of Navigational search

Home page and named page finding tasks are easy to judge, introducing only the minor complication of multiple URLs for the same page. The judging criterion is, "is this the page I wanted ?", i.e. the homepage of the entity I was thinking of or the page I named.

For navigational tasks, success at n documents retrieved (i.e. did the right answer appear in the first n results) or mean reciprocal rank of the first right answer are suitable measures. An even better measure might weight the ranks according to relative prominence within the result list.

For example, the first item in the list is more prominent than the rest; Items "above the fold" (i.e. those which can be seen without scrolling) are more prominent than those below it; Items on the first page of results are more prominent than the second. However, these weightings depend upon the formatting of the results page, the number of results presented per page and the size of the browser window in which the searcher is viewing the results.

Evaluating the effectiveness of Informational Search

Three types of informational search may be identified:

TREC ad hoc Find me a selection of documents relevant to this topic. For example, "I'm writing a paper on this topic and I need background information on all aspects, plus references". The TREC ad hoc methodology is generally applicable apart from presentational issues discussed below and the greater prevalence of duplicate and near-duplicate documents.

A further methodological issue considered by the Track is that of indirect relevance - whether documents which link directly to a relevant page should be considered partly useful. There is no simple answer because the usefulness of a link depends upon how many other outgoing links there are from that page, how prominent the link is on that page, and how easy it would be for a searcher to tell that the link was likely to lead to relevant content.

Topic distillation Make a list of the key resources on some broad topic, similar to those compiled by human editors at DMOZ, Yahoo! or LookSmart.

It is vital to understand, but difficult to explain, how very different this type of search is from that which we have just labelled "TREC ad hoc". Unlike newspaper archives represented in TREC Ad Hoc, the Web (like many other webs):

- is organised into sites, each of which typically provides an entry page giving an overview of the site, navigational links to subsites and a local search capability;
- is extensively interlinked;
- provides services as well as information; and
- can instantaneously generate customised documents in response to requests (such as theatre seating availability).

These characteristics allow webs to be used for different purposes and encourage the use of the "search-and-browse" paradigm in which the searcher types a broad query, gets a list of prominent websites, goes to one or more of the sites and browses (or searches) locally. By visiting the site entry pages the searcher quickly acquires an understanding of how the topic is organised, what vocabulary is used etc.

To illustrate, imagine a prospective Computer Science PhD student has heard of the discipline of Information Retrieval and would like to know more about it: read an introductory textbook, find out who are the leading researchers, list some active departments, know which are the main conferences, and so on.

If we pretended that the Web were a structureless, unlinked, source-anonymised newspaper archive, then the query "information retrieval" would return huge numbers of matching documents (as at 18 May 04, Google estimated 2.7 million, Yahoo! 4.6 million) and the ranking would be based on the similarity of the text to that of the query (essentially the density of references to those terms.) If someone had created a document containing nothing but a hundred repetitions of the phrase "information retrieval", that document would almost certainly rank first. Such a ranking is very unlikely to provide quick access to the sort of understanding the PhD candidate wanted and it is almost certain that a chain of queries would ensue, trying to narrow down to the valuable information expected to be there.

By contrast, rankings produced by the two large-coverage Web search services at the time of writing (Google and Yahoo!) do an impressive job of presenting a list of key resources on the first page: the home page of the van Rijsbergen textbook, an IR bibliography, the home page for CIIR at UMass, a site devoted to the Baeza-Yates & Ribeiro-Neto textbook, the SIGIR information server, the homepage of the Information Retrieval journal, key sites on chemical IR and music IR and an IR research directory on Searchtools.com. One may argue that certain sites deserved to be ranked above others but, for the hypothetical student with their hypothetical requirement, these results are a goldmine at the price of a query so obvious that any such student is able to think of it and no student is too lazy to type it.

Several desirable features of both the Google and Yahoo! rankings for this query are immediately apparent (try it for yourself!):

1. The results are at the right level in the site hierarchy: The CIIR homepage `ciir.cs.umass.edu` is presented ahead of its parent `cs.umass.edu` and its estimated 10,500

children; The van Rijsbergen textbook is represented by it's entry (preface) rather than by arbitrary individual chapters; Hearst's site on the Baeza-Yates & Ribeiro-Neto textbook is represented as `www.sims.berkeley.edu/~hearst/irbook/` rather than `www.sims.berkeley.edu/~hearst/` because the parent site is not restricted to Information Retrieval.

2. No single source (publisher) of information dominates the ranking. Additional results from the same sites are hidden behind a "more results from this site" link.
3. The listed resources are generally well-known and reputable.
4. There is a diversity of resource types.

These positive attributes informed the TREC-2003 methodology for evaluating topic distillation. Key resources were constrained to be site entry pages which:

- were principally devoted to the topic,
- provided credible information on the topic, and
- were not part of a larger site also principally devoted to the topic

No reward was given for diversity among the results returned, but the three criteria adopted prevent credit being given for multiple results from the same source.

Systems were compared on the basis of their ability to retrieve as many key resources as possible early in the ranking. The first ten results were judged and R-Precision (i.e. precision when R documents have been retrieved where R is the number of known key resources) or mean average precision measures were calculated. No reward was given for diversity of resource types and all key resources were judged to have the same value.

Note that finding the appropriate level (within XML elements and sub-elements) for an answer is also a major issue in XML retrieval.

Q&A What is the answer to this question? For example, I'm competing in "Who wants to be a millionaire" and I need help. Q&A tasks have been shown to be well suited to a Web environment (e.g. [37]) but have not yet been evaluated within the Web Track. Readers are referred to [REFER CHAPTER 10] for discussion of methodological issues.

Evaluating the effectiveness of Transactional search

It seems appropriate to evaluate online service finding in similar fashion to informational search, varying only the judging instructions. Measures such as P@n and average precision seem appropriate because it is likely that the searcher wants to see a selection of sites providing the service, in order to be able to compare prices and service details.

Definition of a suitable test collection, including a wide range of e-commerce sites plus useful review and directory pages, would be essential to a meaningful evaluation.

Evaluating the effectiveness of Exhaustive search

Recall-oriented search may be important on the Web or within an enterprise (e.g. Enron) for legal reasons or for creating lists which need to be complete. For example, "find all web pages which mention my name", "find the homepages of every Computer Science department in the United States", "find all the pages which link to the W3C website", "find all pages which contain erroneous JavaScript", or "find every page on the web which makes a claim that eating Brand X hamburgers leads to obesity."

Exhaustive search might seek pages which could also be valid targets of navigational, transactional, or informational search, but often the motivation is different. I might compile a list of all pages where I can buy MP3 players, but not because I want to buy one.

The challenge for future evaluations of Exhaustive search is to find reliable techniques for estimating the full set of matching answers for non-trivial requests. Comparison of systems can be done quite easily but, in exhaustive search, absolute recall is also of interest.

Finding all documents containing the word TREC is an example of a trivial request – the population of relevant documents is easily determined but it is also to be expected that any bug-free retrieval system should be capable of achieving perfect recall within a fixed text collection. (On the Web itself, incomplete recall could result from deficiencies in crawling.) By contrast, finding precedents for a particular legal issue is fraught with difficulty, because no single term can actually capture what is required.

The problem of exhaustive search evaluation, is superficially similar to that of estimating animal populations, where mark-recapture methods [40] can be effective. Unfortunately such methods make assumptions which are not easily satisfied in document retrieval. In particular, relevant documents have unequal probabilities of being "captured" by a particular retrieval run, and exactly the same set of relevant documents will be captured by a subsequent retrieval run with the same parameters.

One possible approach might be to seed the test collection with a set of K known relevant documents chosen in such a way as to be representative of the complete population of possible relevant documents. A retrieval run over the test collection (plus seeds) which retrieved k of the seeds and r other documents judged to be relevant could be used to infer that the number of relevant documents in the original test collection was $R = \frac{rK}{k}$. Multiple heterogeneous retrieval runs could be used to derive and compare a variety of estimates of R.

In estimating populations of relevant documents, there may be value in stratifying by degree of relevance, e.g. on a seven-point scale. Intuitively, estimates of the population of very highly relevant documents may be more reliable than those for the population of those which are peripherally relevant. Furthermore, when rating the usefulness of retrieval tools in exhaustive retrieval applications, recall of highly relevant documents should weigh more heavily.

Note that the density of relevant documents in large collections is too low to permit effective estimation by random sampling.

Presentational issues in Web search evaluation.

The way in which documents are presented to Web Track assessors is far more significant than was the presentation of text-only documents such as newswire reports in the TREC ad hoc task. For example, if a book about "Estimating Animal Populations" were published on the web it might be represented as a number of separate pages: entry page, table of contents, a page for each chapter, and an index. If the entry page consisted only of a scanned image of the front cover with a link to the table of contents, it would be judged irrelevant (in TREC ad hoc) to the topic of "animal populations", if the judge saw only the (empty) text content, even though it is arguably the best entry to the hyperbook.

Ideally, assessors should experience exactly what a real Web searcher would experience if they were carrying out the task in question.

The logistics of presenting pages in a static collection as though they were live pages is fraught with difficulty. If the judging interface allows viewing of images and following of links on the live Web, there is the risk that the images and target pages will have disappeared or changed since the corpus was gathered. On the other hand, if a proxy server is used to serve images and linked pages from within the corpus, many links will be dead because they lead outside the corpus.

Two Web-specific presentational issues are worthy of mention. If one page automatically *redirects* to

another via standard HTTP or HTML mechanisms, both pages should be regarded as correct since what the searcher experiences is more or less the same regardless of which URL they select. This is another source of duplicate content. There are additional difficulties in the case of *framesets*:

- The document which specifies the frameset may have no content at all but may be the appropriate answer to a transactional or informational query because of the useful content in its subsidiary frames (which will be seen by the searcher when the container page is displayed).
- Retrieval systems may retrieve individual frames which are not designed to be displayed in isolation. How should these be displayed for the purpose of judging?

In early VLC and Web evaluations, assessors saw only rendered text of the page being judged. They could not see images or follow links. This may mean that certain pages which would be considered useful in real search are judged useless.

When the .GOV collection was crawled, images and PDF files etc within .gov were saved. However, logistics dictated that they be separated from the text version of the collection and not distributed to participants. This was done because the non-text data size was about four times that of the text, too large to conveniently distribute via CD-ROM. However, CSIRO has recently started distributing test collections on large capacity ATA hard-drives and will make the .GOV images and binaries available on this medium, if required.

The TREC-2002 Interactive Track made use of the .GOV collection via a search engine (Panoptic, `www.panopticsearch.com`) operated by CSIRO. In 2003, the Interactive Track became a subtrack within the Web Track and studied topic distillation tasks within the .GOV collection. This time, the Panoptic search engine was operated by NIST (`ir.nist.gov`) and a mechanism was set up by Ian Soboroff by which images and links were mapped to targets within the original crawl, rather than on the live Web.

VLC/Web Track influence outside TREC

A cursory survey of recent conference proceedings and journal issues reveals that the resources created by the VLC/Web Track are being used quite routinely in studies reported outside TREC. For example, in the years 2000-2002, eight SIGIR papers [34, 2, 44, 15, 32, 3, 4, 39] and four TOIS articles [12, 11, 10, 9] made use of VLC/Web data and several others referred to the track or its methodology. A glance at the same forums for 2003 suggests that usage of VLC/Web Track resources and results is increasing still further.

It is clear that the VLC/Web collections are being used quite widely for tuning, developing and evaluating commercial systems. Examples for which information is available include:

- Copernic Enterprise Search, a commercial search engine specially designed for small and mid-sized enterprises (typically having 5000 to 2 000 000 documents).
- IXE, a C++ class library for indexing and search that is being commercialised by Ideare SpA. The product has been used to build several search services, including online search at www.repubblica.it, one of the major Italian newspapers.
- Microsoft is using Web Track data (among other corpora) to prototype search algorithms which are intended for future product releases.
- Panoptic, CSIRO's metadata-plus-content enterprise search engine with over 30 commercial customers.

- TechRoute Chinese language search engine, which is already used in many orgnizations.

The collections have been distributed to more than 120 organisations world-wide. Many of these groups have not yet participated in TREC.

Limitations of the VLC/Web Track

The VLC/Web Track draws on admirable TREC traditions of building test collections and encouraging group experiments. These are strengths but they also impose limitations:

- It is not feasible to work with collections whose size and link graph complexity even approaches that of the Web. The logistics of distribution are too difficult and too few participants would be able to work with the data. The track can have little to say to Web search engine companies about scalability, efficiency, or large scale graph algorithms.
- The static nature of test collections makes it difficult to use them to explore problems in the areas of crawling or spam rejection.

However, evaluation methodologies applicable to the whole of the Web can be explored in smaller collections. New methods and ideas may also be prototyped and tested on the TREC collections, provided results are not extrapolated to the whole Web without appropriate validation.

Furthermore, enterprise level search (such as modelled .GOV) is commercially and scientifically interesting in its own right. Experimental findings at this scale are applicable in large numbers of search products, potentially affecting a huge number of enterprise/intranet search services.

Resources for ongoing experimentation

Table 2 details the test collections which have been created in the course of the VLC/Web Track. Table 3 details server and connectivity properties for the web collections. Table 4 details the breakdown of the .GOV collection by mime type. Images and the original forms of PDF, msword etc. (around 60gB) were collected and saved. As previously noted they have been used to provide a frozen context for the .GOV collection in the 2003 Interactive experiments.

Likely Future Directions

At the time of writing, guidelines for TREC-2004 Web Track activities are being finalised. Activities are likely to include a mixed Homepage/NamedPage/Topic Distillation task on the existing .GOV collection and some exploratory search tasks within a single-enterprise collection consisting of both websites and email messages.

Enterprise search is economically very important and there is a strong incentive for the Web Track to move in this direction though the challenges are considerable. [1, 17, 21]

Also in TREC-2004, a preliminary version of a new Terabyte Track will revive the tradition of larger scale retrieval within TREC. A new, deeper crawl of the .gov domain has been made, resulting in a collection of approximately 400 gigabytes and 27 million documents. It will be distributed by CSIRO on a single ATA hard drive.

Table 2: Summary of test collections used in VLC and Web Tracks from 1997 onward. CSIRO-distributed collections are accessible via `es.csiro.au/TRECWeb/`.

Collection	Data	#Documents	Ave. Doc. Size	Coll. Size	Availabity	Notes
VLC	Mixed ad hoc and Web	7,492,048	2.8k	20gB	-	No longer distributed.
VLC2 (WT100g)	Web	18,571,671	5.7kB	100gB	CSIRO	From Internet Archive 1997 crawl 1% and 10% samples defi ned & distributed.
WT2g	Web	247,491	8.9kB	2.1gB	CSIRO	Subset of VLC2 (with doc. renaming)
WT10g	Web	1,692,096	6.2kB	10gB	CSIRO	Subset of VLC2 (with doc. renaming)
.GOV	Web	1,247,753	15.2kB	18gB	CSIRO	UWaterloo 2002 crawl of .gov domain. In crawl order. Early termination. Images etc saved.
.GOV2	Web	27M	15kB	400gB	CSIRO	NIST/UWaterloo 2004 more complete crawl of .gov domain. *Under Construction*

Table 3: Link and server statistics for the Web collections. Link density is calculated by dividing the total number of within-collection links (either inlinks or outlinks) by the total number of pages. Cross-server link density is calculated by dividing the total number of within-collection cross-server links (either inlinks or outlinks) by the total number of servers. In the case of WT10g and .GOV connectivity files are distributed on CD-ROM with the data. The same information is available for WT2g from the Web Track website.

Coll.	servers	pages/ server	link density	cross-server link dens.	connectivity data
VLC2	117,101	159			on tape
WT2g			4.71		Website
WT10g	11,680	144	4.77	14.7	on CD-ROM
.GOV	7794	160	8.95	317	on CD-ROM

Table 4: Types of document (mime-type) within the .GOV collection. (Excluding non-text documents.)

text/html	appl/pdf	text/plain	appl/msword	appl/postscript	other
1,053,110	131,333	43,753	13,842	5,673	42

Conclusions

The VLC/Web Track has shown that, in appropriately constituted web search evaluations, retrieval methods based entirely on document content can be substantially outperformed by others which make use of "web evidence", such as anchortext, link measures, and URL or site structure. This has been demonstrated for both informational (topic distillation) and navigational (homepage finding and named page finding) tasks. It is probable that "web evidence" will in the future be found to be similarly important in transactional tasks. It is also possible that page popularity measures (such as click-through data) may be useful in web search.

This conclusion is unsurprising as commercial search companies have known for years that web evidence was invaluable when searching the whole Web. However, the track has shown that the nature of the task is important, and contributed an understanding of the relative merits of many different types of web evidence. Anchor text evidence is highly effective on a range of tasks and in many different webs. Query independent evidence is harder to exploit because of the necessity to combine with query-dependent scores. URL structure is highly effective when websites are static but less so when sites are dynamically generated and URLs have the same form. In enterprise-scale webs where spam is not an issue, simple inlink counts seem to work as well as more sophisticated variants such as PageRank.

The discovery, within the Track, that URL structure could be exploited effectively was a novel contribution.

Another significant outcome of VLC/Web Track (and related) experiments has been to demonstrate that web evidence can strongly contribute to effectiveness even in relatively tiny webs. Significant effectiveness gains due to anchortext and other web evidence have been documented in search tasks over .GOV (1.25 million pages) and even over the artificially constructed WT2g collection (0.25 million pages). Within enterprise webs of only a few thousand pages, anchortext and other web evidence can be used to identify key sites from within large numbers of "relevant" pages.

Many participants and onlookers were shocked when initial Web Track experiments using TREC Ad Hoc methodology showed no benefit whatever from the use of web evidence. In hindsight, the explanation is breathtakingly simple – the evaluation methodology did not accurately represent the major phenomenon (typical web search) we were purporting to study.

The reality is that many web searchers regard entry pages of authoritative relevant sites as more valuable than are isolated pieces of relevant text. They have come to expect that entry pages of key relevant sites will appear at the top of search engine rankings.

An effective web search tool must be able to support the search-and-browse paradigm by bringing the most valuable matching resources (frequently site homepages) to the top, while preventing the list from being flooded by pages from a single source. Our early evaluations failed to recognize this, and we scored individual pages on their text content only, using only a binary scale. For searchers prepared to treat the Web as a newspaper archive and scan thousands of relevant documents, this was acceptable, but on the Web such people are a rare breed.

More recent evaluation methodologies adopted by the track, such as home page and named page

finding and the 2003 Topic Distillation task, have rectified the methodology problem by considering only the most valuable web sites (or specific named pages) on a topic. This approach automatically penalises multiple results from the same source and gives infinitely higher value to the most valuable resources than to individual relevant documents.

A more subtle evaluation methodology seems possible in which retrieved documents are given scores (across a wide range, not just a few degrees of text match) commensurate with the value they have to the prototypical Web searcher. The formulation of such a scoring function would have to address the issues of duplicate documents and source diversity and would probably score zero for duplicates and fractions for subsequent documents from the same source. Such a methodology could largely avoid the need for segregating tasks into search types (e.g. topic relevance, homepage finding, topic distillation) as topics for which there were no key websites or other extra valuable resources would automatically fall back to evaluation in standard TREC ad hoc fashion.

Acknowledgements

The late Paul Thistlewaite was a driving force behind the VLC/Web Track and Donna Harman, Ellen Voorhees and Ian Soboroff at NIST have been very supportive of it. The many individual donors of data to the VLC collection have been thanked in the appropriate TREC Proceedings but are acknowledged again here. John Ritchie and Mark Sanderson, then of the University of Glasgow provided great assistance in negotiating for permission to use some of this data. John O'Callaghan and Darrell Williamson showed vision in agreeing to let ACSys distribute data sets, as did Murray Cameron in agreeing that CSIRO should continue the work. A large number of relevance assessors from ACSys, CSIRO and NIST were employed over the years and their work was indispensible.

The willingness of Brewster Kahle of the Internet Archive to supply what was at the time a huge crawl was instrumental in the success of the track. Edward King (CSIRO) provided data conversion facilities and Peter Bailey (then of ANU) did a great job in planning and creating WT10g. Edwin Cooper (the Electric Monk) and Michael Moricz (Alta Vista) provided query sets which formed the basis of many evaluations. Charlie Clarke (UWaterloo), Ed Fox (Virginia Tech) and Ian Soboroff made possible the .GOV collection.

Many ACSys administrative staff assisted in the distribution of test collections and since CSIRO took on the role, all aspects of distribution have been capably looked after by Daphne Bruce.

We are very grateful for the cooperation and assistance of the very large number of people who have helped build the infrastructure of the track and to support its operation. We also acknowledge the many groups who have participated in VLC/Web Track evaluations and thereby contributed to the advancement of knowledge in this important area.

References

[1] Mani Abrol, Neil Latarche, Uma Mahadevan, Jianchang Mao, Rajat Mukherjee, Prabhakar Raghavan, Michel Tourn, John Wang, and Grace Zhang. Navigating large-scale semi-structured data in business portals. In *Proceedings of the 27th VLDB Conference*, pages 663–666, Roma, Italy, 2001. `www.vldb.org/conf/2001/P663.pdf`.

[2] Vo Ngoc Anh, Owen de Kretser, and Alistair Moffat. Vector-space ranking with effective early termination. In *Proceedings of ACM SIGIR 2001*, pages 35–42, New Orleans, LA, 2001.

[3] Vo Ngoc Anh and Alistair Moffat. Impact transformation: effective and efficient web retrieval. In *Proceedings of ACM SIGIR 2002*, pages 3–10. ACM Press, 2002.

[4] Dirk Bahle, Hugh E. Williams, and Justin Zobel. Efficient phrase querying with an auxiliary index. In *Proceedings of ACM SIGIR 2002*, pages 215–221. ACM Press, 2002.

[5] Peter Bailey, Nick Craswell, and David Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing and Management*, 39(6):853–871, 2003. `www.ted.cmis.csiro.au/~dave/cwc.ps.gz`.

[6] Tim Berners-Lee. List of WWW servers, 1992. `www.w3.org/History/19921103-hypertext/hypertext/DataSources/WWW/Servers%.html`.

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

[8] Robert Calliau. A little history of the World Wide Web, 1995. `www.w3.org/History.html`.

[9] Adam Cannane and Hugh E. Williams. A general-purpose compression scheme for large collections. *ACM Transactions on Information Systems (TOIS)*, 20(3):329–355, 2002.

[10] Claudio Carpineto, Giovanni Romano, and Vittorio Giannini. Improving retrieval feedback with multiple term-ranking function combination. *ACM Transactions on Information Systems (TOIS)*, 20(3):259–290, 2002.

[11] Abdur Chowdhury, Ophir Frieder, David Grossman, and Mary Catherine McCabe. Collection statistics for fast duplicate document detection. *ACM Transactions on Information Systems (TOIS)*, 20(2):171–191, 2002.

[12] Charles L.A. Clarke and Gordon V. Cormack. Shortest-substring retrieval and ranking. *ACM Transactions on Information Systems*, 18(1), 44-78 2000.

[13] Kevyn Collins-Thompson03, Paul Ogilvie, Yi Zhang, and Jamie Callan. Information filtering, novelty detection and named-page finding. In *Proceedings of TREC-2002*, pages 107–118, Gaithersburg, November 2002.

[14] Nick Craswell, Peter Bailey, and David Hawking. Is it fair to evaluate web systems using trec ad hoc methods? ACM SIGIR '99 Workshop on Web Retrieval, 1999. `pastime.anu.edu.au/nick/pubs/sigir99ws.ps.gz`.

[15] 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`.

[16] Nick Craswell, David Hawking, Ross Wilkinson, and Mingfang Wu. Overview of the trec-2003 web track. In *Proceedings of TREC 2003*, Gaithersburg, Maryland USA, November 2003. `trec.nist.gov`.

[17] Ronald Fagin, Ravi Kumar, Kevin S. McCurley, Jasmine Novak, D. Sivakumar, John A. Tomlin, and David P. Williamson. Searching the workplace web. In *Proceedings of WWW2003*, Budapest, Hungary, May 2003. `www2003.org/cdrom/papers/refereed/p641/xhtml/p641-mccurley.html`.

[18] Michael Gordon and Praveen Pathak. Finding information on the World Wide Web: The retrieval effectiveness of search engines. *Information Processing and Management*, 35(2):141–180, March 1999.

[19] D. K. Harman, editor. *Proceedings of TREC-1*, November 1992. NIST special publication 500-207.

[20] David Hawking. Overview of the TREC-9 Web Track. In *Proceedings of TREC-9*, Gaithersburg MD, November 2000. NIST special publication 500-249, `trec.nist.gov`.

[21] David Hawking. Challenges in enterprise search. In *Proceedings of the Australasian Databases Conference ADC2004*, Dunedin, New Zealand, January 2004.

[22] 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`.

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

[24] David Hawking, Nick Craswell, Peter Bailey, and Kathleen Griffiths. Measuring the quality of public search engines, 2000. `pastime.anu.edu.au/TAR/Search_Engines_Conf/`.

[25] David Hawking, Nick Craswell, Peter Bailey, and Kathleen Griffiths. Measuring search engine quality. *Information Retrieval*, 4(1):33–59, 2001. preprint at `www.ted.cmis.csiro.au/~dave/INRT83-00.ps.gz`.

[26] David Hawking, Nick Craswell, and Paul Thistlewaite. Overview of TREC-7 Very Large Collection Track. In *Proceedings of TREC-7*, pages 91–104, November 1998. NIST special publication 500-242, `trec.nist.gov/pubs/trec7/t7_proceedings.html`.

[27] David Hawking, Nick Craswell, Paul Thistlewaite, and Donna Harman. Results and challenges in web search evaluation. *Proceedings of WWW8*, 31:1321–1330, 1999. `www8.org/w8-papers/2c-search-discover/results/results.html`.

[28] David Hawking and Stephen Robertson. On collection size and retrieval effectiveness. *Information Retrieval*, 6(1):99–150, 2003.

[29] David Hawking and Paul Thistlewaite. Overview of TREC-6 Very Large Collection Track. In *Proceedings of TREC-6*, pages 93–106, November 1997. NIST special publication 500-240, `trec.nist.gov`.

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

[31] Internet Archive. Building a digital library for the future, August 1997. `www.archive.org/`.

[32] Wessel Kraaij, Thijs Westerveld, and Djoerd Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings of ACM SIGIR 2002*, pages 27–34, Tampere, Finland, 2002.

[33] Yves Lafon. HTTP - Hypertext Transfer Protocol. `www.w3.org/Protocols/`, accessed Sep 2002.

[34] Zihong Lu and Kathryn S. McKinley. Partial replica selection based on relevance for information retrieval. In *Proceedings of ACM SIGIR 2000*, pages 248–255, Athens, Greece, 2000.

[35] Clifford A. Lynch. The z39.50 information retrieval standard, part 1. *D-Lib Magazine*, April 1997. `mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/april97/04lynch.html`.

[36] NIST. Trec overview. `trec.nist.gov/overview.html`, accessed September, 2002.

[37] Dragomir R. Radev, Kelsey Libner, and Weiguo Fan. Getting answers to natural language questions on the web. *JASIST*, 53(5), 2002. `http://www.asis.org/Publications/JASIS/vol53n05.html`.

[38] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[39] Falk Scholer, Hugh E. Williams, John Yiannis, and Justin Zobel. Compression of inverted indexes for fast query evaluation. In *Proceedings of ACM SIGIR 2002*, pages 222–229. ACM Press, 2002.

[40] G.A.F. Seber. *The Estimation of Animal Abundance and Related Parameters*. Charles Griffin & Co., London, 2nd edition, 1982.

[41] 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`.

[42] 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.

[43] RFC1738: Uniform resource locators (URL). `www.w3.org/Addressing/rfc1738.txt`. Accessed 25 Sep 2001.

[44] Ellen Voorhees. Evaluation by highly relevant documents. In *Proceedings of SIGIR 2001*, pages 74–82, New Orleans, LA, 2001.

[45] Thijs Westerveld, Wessel Kraaij, and Djoerd Hiemstra. Retrieving web pages using content, links, urls and anchors. In *Proceedings of TREC-2001*, pages 663–672, Gaithersburg, MD, Nov 2001. NIST. `trec.nist.gov`.

[46] Robert Zakon. Hobbes' internet timeline, 2002. `www.zakon.org/robert/internet/timeline/`.

[47] Min Zhang, Ruihua Song, Chuan Lin, Shaoping Ma, Zhe Jiang, Yijiang Jin, Yiqun Liu, and Le Zhao. Thu trec2002 web experiments. In *Proceedings of TREC-2002*, pages 591–594, Gaithersburg, November 2002.