Overview of the TREC-8 Web Track

David Hawking*
CSIRO Mathematical and Information Sciences,
Canberra, Australia
Ellen Voorhees
National Institute of Standards and Technology
Gaithersburg MD, USA
Nick Craswell and Peter Bailey
Department of Computer Science, ANU
Canberra, Australia
David.Hawking@cmis.csiro.au, Ellen.Voorhees@nist.gov,{nick,peterb}@cs.anu.edu.au

February 22, 2000

Abstract

The TREC-8 Web Track defined ad hoc retrieval tasks over the 100 gigabyte VLC2 collection (Large Web Task) and a selected 2 gigabyte subset known as WT2g (Small Web Task). Here, the guidelines and resources for both tasks are described and results presented and analysed.

Performance on the Small Web was strongly correlated with performance on the regular TREC Ad Hoc task. Little benefit was derived from the use of link-based methods, for standard TREC measures on the WT2g collection. The number of inter-server links within WT2g may have been too small or it may be that link-based methods would have worked better with different types of query and/or with different types of relevance judgment. In fact, a small number of link-based runs proved to be much more effective than their content-only baseline at finding documents which linked to documents judged relevant.

A variety of issues were investigated by participants in the Large Web Task. One group investigated the use of PageRank scores and found no benefit on standard TREC measures. Engineering improvements by several groups led to either considerable reduction in query processing time or reduction in the amount of hardware necessary to maintain comparable performance.

1 Introduction

The TREC-8 Web Track activities centred on two tasks: the Small and the Large Web Tasks. The latter featured the 100 gigabyte, 18.5 million webpage VLC2 collection described in last year's VLC Track overview [Hawking et al. 1998] and on the Web Track website [ACSys]. The former made use of a 2 gigabyte, 250,000 document subset of the VLC2, distributed on CD-ROM as the WT2g collection. Note that documents in WT2g are given different document numbers to the ones they had in the VLC2 (to enable easy extraction of the document) but include the original document numbers within `DOCOLDNO` tags.

As it turned out, the Large and Small Web sub-tracks had very little in common apart from the use of spidered Web data. Accordingly, they will be described separately.

*The authors wish to acknowledge that this work was carried out partly within the Cooperative Research Centre for Advanced Computational Systems established under the Australian Government's Cooperative Research Centres Program.

2 Small Web Task

The focus of the Small Web Task was on answering two specific questions:

1. Do the best methods in the TREC Ad Hoc task also work best on the WT2g collection of Web data?
2. Can link information in Web data be used to obtain more effective search rankings than can be obtained using page content alone?

2.1 Topics and assessments

The Small Web Task used the TREC-8 Ad Hoc topics. Submissions were judged by NIST using the same tools and presentation as used for the Ad Hoc documents. The small web and ad hoc documents for a given topic were judged by the same assessor, almost always the topic author. There was a single pool of documents for each topic, but since pools were sorted alphabetically by DOCNO, all small web track documents appeared after the ad hoc documents. (Assessors were free to jump around in the pool when judging, but seldom did so.)

A public-domain HTML to ASCII converter, substantially modified by Ellen Voorhees, was used to render the entire web collection into ASCII. The rendered version is what the assessors judged. The rendering threw away all images, scripts, and frames, replacing them with a simple notice such as `[IMAGE GOES HERE]`, unless the HTML provided ALT text in which case that was used (this occurred very rarely). The text of tables was retained, and a rough approximation of its formatting. Links were NOT rendered. This rendered collection is what was indexed to enable the assessor to do PRISE [NIST] searches during topic development. However, wholesale pre-editing of the source was needed to eliminate Word and PowerPoint documents, Chinese, Japanese, and other non-text data.

2.2 Judging pools

The number of runs judged was 27, giving a maximum pool size of 2700. The mean actual pool size was 950, 35.2% of the maximum, while the mean number of relevant documents over the 50 topics is 45, which is 4.8% of the number of documents judged.

41.2% of the documents in the pool were contributed by both a content-only and a content-link run. 39% of the pool was contributed by only content-only runs, and 19.8% by only content-link runs. The statistics for the relevant documents are more skewed: 76.3% of the relevant documents were found by runs of both types, 19.2% of the relevant documents by only content-only runs, and 4.5% of the relevant documents by only content-link runs.

Each of the 17 groups that submitted Small Web Task runs found some relevant documents that no other group retrieved in the top 100 ("unique relevants"). The largest totals over the 50 topics for unique relevants are 89 for Rutgers(Davison), 60 each for Claritech and RMIT, and 36 for IRIT. The totals for the other groups ranged from 5 to 30.

The pool statistics for the Small Web Task are roughly comparable to the statistics for the main Ad Hoc Task. For the Ad Hoc pools, the mean actual pool size was 1736 out of 7100 possible (24.5%) and the mean number of relevant documents is 94 (5.4% of what was judged). The number of unique relevants was comparatively larger, mostly reflecting the use of manual runs: the top three totals for unique relevants are 478 for MITI, 114 for Oracle, and 80 for IIT.

2.3 Definition of the WT2g dataset

In order to address the Small Web questions, a subset of the 100 gB VLC2 collection was needed which:

- was comparable in size to the TREC Ad Hoc collection (so as not to discourage participation, and to avoid perturbing collection parameters more than necessary);

- was likely to contain a reasonable quantity of material relevant to the TREC-8 Ad Hoc topics;
- included naturally defined sub-collections; and
- contained an interesting quantity of closed hyperlinks (with both source and target page within the subcollection).

The second and third requirements ruled out a uniform 2% sample.

The method of choosing the WT2g subset collection was entirely heuristic. We started by identifying all the distinct hosts represented in the 100 gigabyte collection. Then we counted how many relevant documents were found in the VLC tasks (using TREC-7 ad hoc topics) and ranked the hosts in order of decreasing relevant document density. Finally, we collected together all the documents from the top-ranked hosts until we reached a little over 2 gigabytes of data. The number of hosts represented is 956.

We expected that:

- the much higher density of relevant documents on TREC-7 topics than the average for the VLC2 would lead to a similarly higher density for TREC-8 topics; and
- because all available documents from each host were included, the proportion of dead links in this 2 gigabyte sample would be much less than for a randomly chosen sample of the same size.

Table 1: The density of known relevant documents in VLC2, WTg and the current TREC Ad Hoc collections for TREC-7 and TREC-8 Ad Hoc topics. The original T7 judgments for WT2g were obtained from runs against the whole VLC2 collection and were understood to be incomplete. After TREC-7, ACSys judged some additional documents within WT2g for the T7 Ad Hoc topics. These are reported in the line marked "T7+new".

Judgments	Collection	Density of relevant docs
T7	VLC2	6482/18571671 = 0.03%
T7	WT2g	3105/247491 = 1.25%
T7+new	WT2g	6495/247491 = 2.62%
T7	Ad Hoc	4674/528155 = 0.89%
T8	WT2g	2279/247491 = 0.92%
T8	Ad Hoc	4728/528155 = 0.90%

Table1 and Figure 1 show the densities of known relevant documents for the TREC-7 and TREC-8 Ad Hoc topics within various collections. Naturally, there may be considerable variation from one topic to another.

NIST assessors referred to the WT2g collection during the process of ad hoc topic generation. The assessors checked the number of relevant documents in the Web collection once they had a candidate topic from searching the ad hoc collection. The procedure was the same for both collections:

1. Use PRISE [NIST] to retrieve 25 documents.
2. If, after judging 25 documents there are at least 1 and less than 20 relevant documents, perform feedback and judge the resulting top 100; Otherwise stop since candidate topic is not acceptable.

There was a nominal cut-off of at least 10 relevant documents in each collection to be selected as a final topic, but slightly less than that were accepted on some topics.

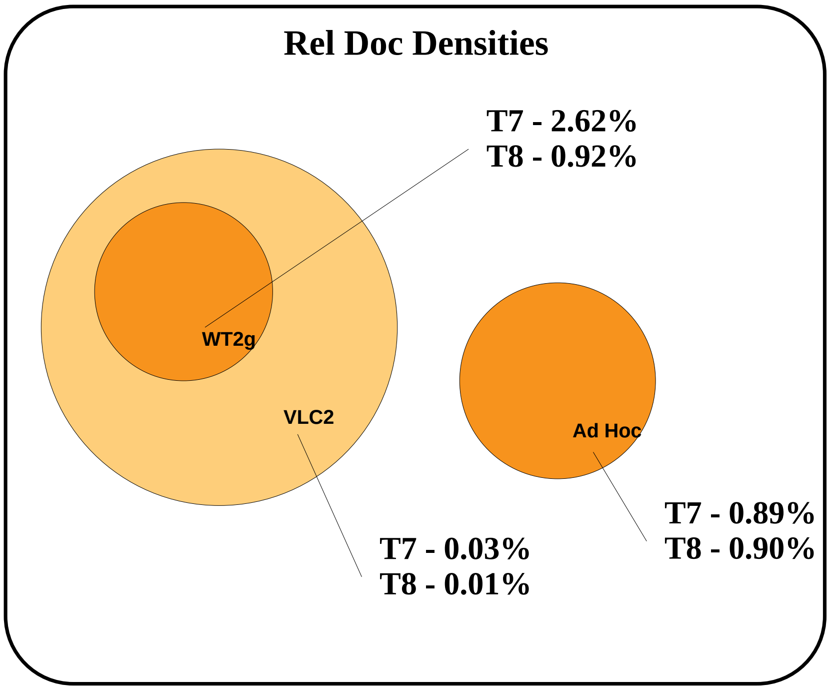

Figure 1: Pictorial representation of the data in Table 1.

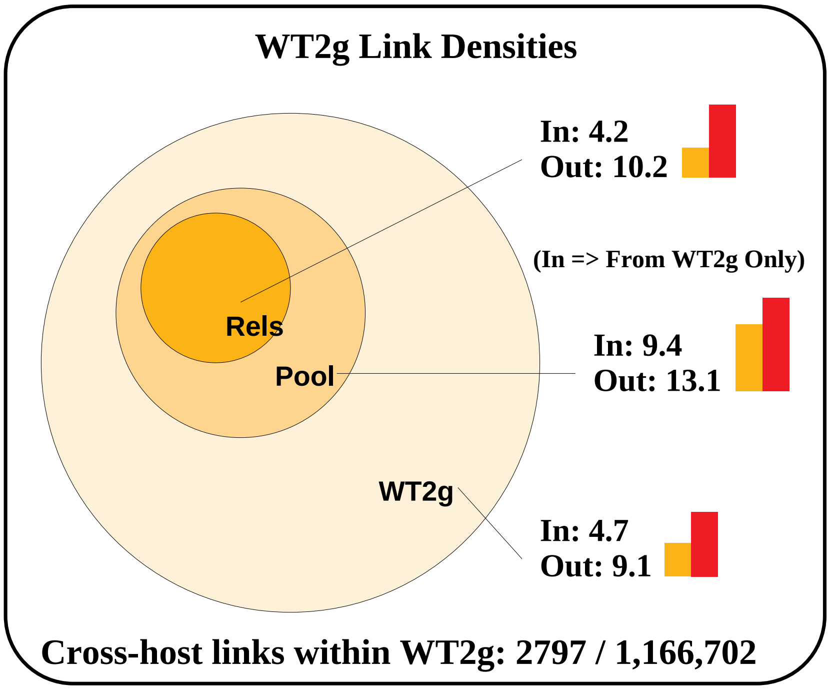

Figure 2: Pictorial representation of the data in Table 2.

2.4 Connectivity data

Nick Craswell developed software for extracting hyper-link connectivity information from WT2g. ACSys made that data available in two ways. First a connectivity server was made available on the Web. For each input URL the server would respond with a list of incoming links from other WT2g documents and outgoing links. However, because of network latencies and the extra client-side coding needed to resolve each URL to a canonical form and quote any special characters to avoid confusing the CGI script, the connectivity server was difficult to use. Accordingly, the connectivity data was also distributed by `ftp` in a highly compressed format based on WT2g document numbers. The out-links file consisted of, for each document d, the document numbers of the documents d links to. The in-links file was similar, but listed for each d the document numbers of documents linking to d.

2.5 Interconnectedness of WT2g

Table 2: The average number of outgoing links per document for various source sets, broken down by destination. The number of documents in WT2g is 247,491, the number in the assessment pool is 35,089 and the number in the relevant set is 2279.

Link source	Link target	Total Links	Links per source doc	Links per target doc
WT2g	Universe	2,259,952	9.13	-
WT2g	WT2g	1,166,702	4.71	4.71
WT2g	Assessment pool	330,295	1.33	9.41
WT2g	Relevant set	9512	0.04	4.17
Assessment pool	Universe	460,449	13.12	-
Assessment pool	WT2g	217,288	6.19	0.88
Assessment pool	Assessment Pool	88,468	2.52	2.52
Assessment pool	Relevant set	3579	0.10	1.57
Relevant set	Universe	23,337	10.24	-
Relevant set	WT2g	10,843	4.76	0.04
Relevant set	Assessment Pool	5181	2.27	0.15
Relevant set	Relevant set	711	0.31	0.31

Table 2 and Figure 2 show the density of links between different sets of documents. On average, each document within the collection includes 9.13 outgoing links. 52% of these links reference another document within WT2g but only 0.12% reference a different server within WT2g. It is not known at this stage, what proportion of the dead links (those whose target lies outside WT2g) are inter-server links and how many are references to same-server pages which happen to be missing from the VLC2[1].

2.6 Summary of participation

Seventeen groups submitted a total of 44 runs, 24 content-only and 20 making use of links.

2.7 Content-only runs

Groups which submitted exactly corresponding runs in the Ad Hoc and Small Web Tasks were asked to supply runids and evaluation results for the Ad Hoc task. The corresponding average precision scores for these runs on the two tasks are tabulated in Table 4 and plotted against each other in Figure 3. It should be noted that some of the pairs of runs did not exactly correspond. AT&T used duplicate elimination

[1] Note that if any pages from a server are included in the WT2g, all VLC2 pages from that server are also included.

Table 3: The best performing content-only runs for each of the 17 participating groups, presented in order of decreasing average precision.

Group	Run tag	Ave. prec.	P@20
Microsoft	ok8wmx	0.3829	0.4520
Fujitsu	Flab8wtdnN	0.3405	0.4010
UMass	INQ620	0.3327	0.4130
MDS/RMIT	mds08w1	0.3220	0.3860
UNeuchatel	UniNEW2Ct	0.3150	0.3940
AT&T	att99wtde	0.3113	0.4110
UWaterloo	uwmt8w0	0.3066	0.3620
ACSys	acsys8wm	0.3009	0.3870
Claritech	CL99WebM	0.2889	0.2880
Illinois	iit99wt1	0.2265	0.3150
Dublin City Uni	DCU99C01	0.1936	0.2510
Seoul Uni	Scai8Web1	0.1854	0.2660
IRIT, Toulouse	Mer8Wctd	0.1638	0.2430
Rutgers	disco2	0.1023	0.1270
Oslo	hio1	0.0927	0.1420
UIowa	uiowaweb1	0.0747	0.1450
UNC	isw50t	0.0291	0.0830

when controlling feedback on the Web runs but not in Ad Hoc. IRIT results constitute the most obvious outlier, but it is not yet clear why.

2.8 Exploitation of links

Table 5 summarises the methods used by the Small Web participants.

Tables 6 – 8 summarise the average precision, P@20 and total-relevant-documents-retrieved scores for each group which submitted at least one content-plus-link run. Each line in these tables gives the baseline performance in Column 2 and the corresponding performance for each link run in the remaining columns. Unfortunately, it is not completely clear that the only difference between the content-plus-link runs and the baseline is the use of links.

The differences between content-plus-link runs and the corresponding baseline are mostly very small and usually negative. The few large differences were all negative.

2.9 Duplicate elimination

Participants were not encouraged to apply duplicate elimination to their runs. It would thus be unfair to penalise runs which included duplicates within their rankings.

Despite a claim that there is at least one topic for which all the relevant documents are clones of each other, it is unlikely that the presence of duplicates would distort relative performance significantly. This is because of averaging over 50 topics and because the presence of irrelevant near-duplicates can degrade performance.

Future Web tracks may adopt evaluation measures which do not reward the presentation of multiple "near-duplicate" pages. However, the following issues need to be resolved:

- What constitutes a near-duplicate?

Table 4: Pairs of corresponding runs in Ad Hoc and Small Web Tasks.

Ad Hoc		Small Web	
Run tag	Ave. prec.	Run tag	Ave. prec.
Scai8Adhoc	.1461	Scai8Web1	.1854
acsys8amn	.2353	acsys8wm	.3009
ok8amxc	.3169	ok8wmx	.3829
att99atdc	.3089	att99wtdc	.3091
att99atde	.3165	att99wtde	.3113
INQ603	.2659	INQ620	.3327
unofficial	.2293	mds08w1	.3220
Mer8Adtd1	.2231	MerWctd	.1638
uwmt8a0	.2143	uwmt8w0	.3066
isa50t	.027	isw50t	.029

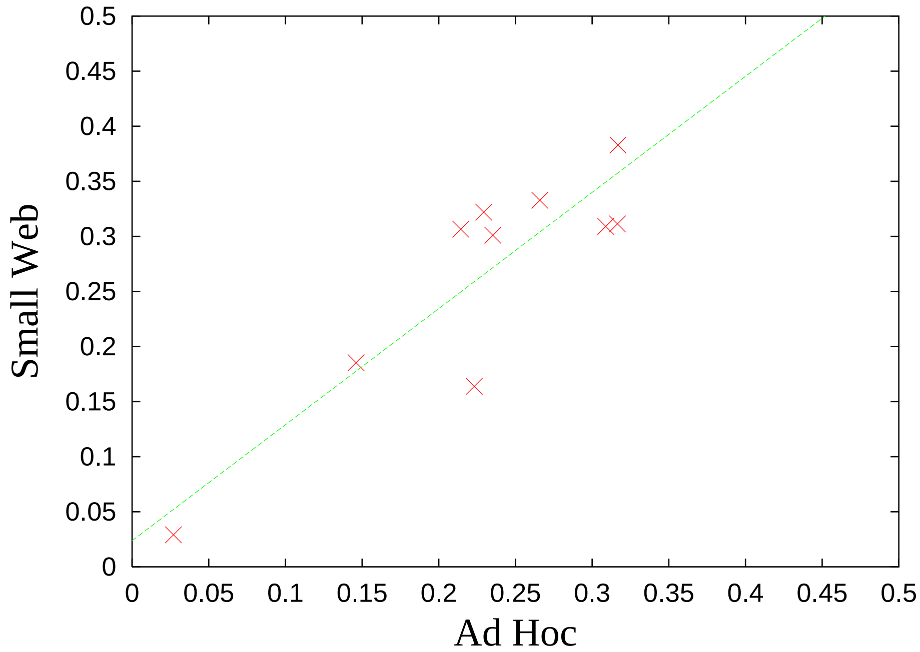

Figure 3: Average precision on the Small Web Task plotted against average precisions on the Ad Hoc task for pairs of runs believed to correspond closely, as per Table 4. Also shown is the line of best (least-squares) fit. The Pearson R coefficient of correlation is 0.884, which is significant at the 0.05 level (two-tailed).

Table 5: Link exploitation methods used by groups participating in the Small Web Task.

Group	Methods
MDS/RMIT	Sibling pages
UniNeuchatel	Kleinberg, PageRank, Spread. Act., PAS
ACSys	PageRank
IIT	Modified Kleinberg
DCU	Inlink/Outlink frequencies
Seoul Nat. Uni.	Score propagation along inlinks
IRIT	Spread. Act.
Rutgers	like Kleinberg
Oslo College	?
UIowa	?
Claritech	Kleinberg

Table 6: Comparison of average precision scores for runs using links with those of the corresponding baseline runs. Four link-based runs out of the 20 submitted achieved scores which were slightly higher (numerically) than their baselines. They are highlighted in boldface.

Group	baseline	links1	links2	links3
MDS/RMIT	0.3220	0.3047	0.2878	
UniNeuchatel	0.3150	0.3137		
UniNeuchatel	0.2739	**0.2747**		
ACSys	0.3009	0.3007	0.3007	0.2804
IIT	0.2265	0.2265	0.2264	
DCU	0.1936	**0.1939**	0.1921	
Seoul	0.1854	0.1819		
IRIT	0.1638	0.1488	0.1435	0.1401
Rutgers	0.1023	**0.1087**		
Oslo	0.0927	**0.0972**	0.0945	0.0859
UIowa	0.0747	0.0246		

Table 7: Comparison of P@20 scores for runs using links with those of the corresponding baseline runs. Four link-based runs out of the 20 submitted achieved scores which were slightly higher (numerically) than their baselines. They are highlighted in boldface.

Group	baseline	links1	links2	links3
MDS/RMIT	0.3860	0.3330	0.3590	
UniNeuchatel	0.3940	0.3940		
UniNeuchatel	0.3650	**0.3690**		
ACSys	0.3870	0.3870	0.3700	0.3870
IIT	0.3150	0.3150	0.3150	
DCU	0.2510	0.2490	0.2510	
Seoul	0.2660	0.2660		
IRIT	0.2430	0.1950	0.2160	0.2130
Rutgers	0.1270	0.1110		
Oslo	0.1420	**0.1670**	**0.1580**	**0.1430**
UIowa	0.1450	0.0290		

Table 8: Comparison of total relevant documents retrieved across all 50 topics, for runs using links with those of the corresponding baseline runs. The link-based runs which retrieved more relevant documents than their baselines are shown in bold.

Group	baseline	links1	links2	links3
MDS/RMIT	1872	1872	**1878**	
UniNeuchatel	1880	1869		
UniNeuchatel	1796	1795		
ACSys	1835	1834	1748	1834
IIT	1575	1572	1568	
DCU	1017	1017	1017	
Seoul	1500	**1504**		
IRIT	1286	**1338**	1258	**1352**
Rutgers	1041	**1072**		
Oslo	1292	1288	**1394**	1176
UIowa	1074	1074		

Table 9: Number of directly and indirectly relevant documents found in the WT2g collection. Note that documents may count more than once – once for each topic for which they are relevant.

Type	Number
Directly relevant	2279
Directly OR indirectly relevant	8838
Directly AND indirectly relevant	242
Indirectly but not directly relevant	6559

- How to score near-duplicates within a ranking. Zero for all near-duplicates after the first? Or, fractional scores?

2.10 Scoring taking into account links

It is possible that a link-based retrieval method may return a significant number of documents which, although they contain little or no relevant content, contain links to relevant documents. Such indirectly relevant documents are of value to a searcher in a Web search context because they provide a low-cost path to relevant documents.

Table 10: Additional relevant documents found when documents which directly link to relevant documents are considered to be indirectly relevant. The left part of the table considers documents found in the top 20 rankings and the right part considers documents found anywhere within the top 1000 results. In each part, the base column shows the total indirectly relevant documents found across all 50 topics. The number of indirectly relevant documents found by the link-based runs is shown relative to the number found by the corresponding baseline run. If every indirectly relevant document were considered to have the same weight as a directly relevant one, each indirectly relevant document found in the top 20 would add 0.001 to the original precision @ 20.

	top 20				top 1000			
Group	base	links1	links2	links3	base	links1	links2	links3
MDS/RMIT	17	+6	+3		525	0	+85	
UniNeuchatel	13	0			597	+838		
UniNeuchatel	13	0			590	+772		
ACSys	22	0	+1	0	549	-1	-53	-1
IIT	18	0	0		476	+344	+254	
DCU	12	0	0		137	0	0	
Seoul	16	+2			418	+37		
IRIT	17	+13	+10	+6	438	+96	-13	+17
Rutgers	11	+82			283	+239		
Oslo	32	+1	+2	-14	516	+16	+34	-62
UIowa	25	+17			394	0		

The benefits of link-based retrieval may thus be underestimated because `trec_eval` does not take this into account. ACSys has attempted to determine whether this was the case by looking at the indirectly relevant documents retrieved by the runs listed in Table 7.

The WT2g connectivity data (see `http://pastime.anu.edu.au/WAR/WT2g_Links/ilink_WTonly.gz` and the Small Web qrels file were used to find the set of documents which link directly to relevant documents. Table 9 gives the numbers of directly and indirectly relevant documents.

Table 10 reports the number of indirectly but not directly relevant documents included in the runs listed in Table 7, both in the full (top 1000) rankings and in the top 20 rankings. The results are presented so as to highlight any differential tendency of link-based runs to find indirectly relevant documents.

Considering the "top 1000" part of the table, several link-based runs show differentially higher retrieval of indirectly relevant documents. If all indirectly relevant documents are considered to be as valuable as directly relevant ones, the total set of relevant documents nearly quadruples in size, recall values for the top 1000 rankings declines sharply and the ordering of several (content, content+link) pairs changes. Most notable of these changes are those of the University of Neuchatel whose content+link runs out-recall their corresponding baselines by 33% (2477 v. 3304) and 32% (2386 v. 3157) and IIT whose content+link runs now out-recall the baseline by up to 17%. These comparisons are made on the basis of total number of relevant documents retrieved over the 50 topics.

Inspection of the table suggests that link methods used by the University of Neuchatel, by IIT and in one of the IRIT runs resulted in differentially greater retrieval of indirectly relevant documents to an extent that might possibly change the ranking of corresponding pairs of runs on recall.

Considering the "top 20" part of Table 10, the Rutgers link-based run has a much higher differential retrieval of indirectly relevant documents than any other link-based run. If every indirectly relevant document were accorded the same weight as a directly relevant one, the Rutgers baseline P@20 would increase to 0.138 (from 0.127) and the link-based run to 0.204 (from 0.110). This appears to be the only pair of runs for which the consideration of indirectly relevant pages may change the ranking of the runs on P@20. Furthermore, the benefit implied by these figures is almost certainly overstated, due to the assumption of equal worth for directly and indirectly relevant pages.

The worth of an indirectly relevant document to a searcher depends upon how easy it is to find the link to the directly relevant page(s). This is influenced by page layout factors not easily determinable automatically. For example:

1. whether the visual rendition of the link attracts attention;
2. whether the link is at the top of the document or in some other prominent position;
3. whether the anchor and context of the link allow the searcher to identify that the target of the link is likely to be relevant;
4. whether there are other similarly attractive links which, in fact, lead to irrelevant pages.

On average, the value of an indirectly relevant page is likely to be considerably less than that of a directly relevant page. Accordingly, scores on TREC measures, revised to take into account indirect relevance, have not been presented because they would depend upon an arbitrary assignment of relative weight for indirectly relevant pages.

2.11 Small Web Task discussion and conclusions

The University of Neuchatel and Fujitsu Laboratories report that they could find no correlation between relevance on the TREC-7 topics and link-based measures.

It seems fairly clear that, in this year's Small Web Task, no measurable benefit was gained on standard TREC retrieval measures through use of links. A small number of link-based runs benefited substantially on recall, and one on P@20 provided that indirectly relevant documents are assigned the same value as directly relevant ones.

The following questions arise:

1. Is the WT2g collection big enough and does it include enough links to permit effective operation of the link-based methods? Figure 2 shows that there are in fact a lot of links, but only a very small number of cross-server links.
2. Are link-based methods more likely to be effective for types of information need other than those modelled by TREC Ad Hoc topics. For example, locating a library's on-line catalogue or the home page of a particular travel agent.
3. Would link-based methods seem more effective if the TREC relevance assessment model were expanded to recognise that some pages are much more valuable than those which are merely relevant. For example, the desired on-line catalogue page may be of far more use to the searcher than learned papers about library cataloguing systems.

Table 11: Summary of Results for all submitted runs over the full 100 gigabyte collection. Note that the UNC runs were submitted after the deadline and include many unjudged documents. Note also that the query processing times reported for Fujitsu correspond to the case where queries were processed as a sequential batch. Fujitsu achieved better times on the same hardware using two processes: 0.40 sec (`fl8wlnsb`); 0.75 sec (`fl8wlnsr`); 0.39 sec (`fl8wlsb`). See the Fujitsu paper for details.

Group	Runid	Cost (k$)	idx_time (hr.)	idx_size (gB)	qp_time (sec.)	p20
ACSys	acsys8lw0	7	8.48	5.78	3.74	0.3360
ACSys	acsys8lw0_pr1	7	104	6.46	3.91	0.3360
ACSys	acsys8lw0_pr10	7	104	6.46	3.87	0.3350
AT&T	att99vlci	115	8.62	23.9	0.516	0.55650
AT&T	att99vlcm	115	8.62	23.9	0.516	0.5470
Fujitsu	fl8wlnsb	41	504	5.10	0.75	0.5100
Fujitsu	fl8wlnsr	41	504	5.10	1.16	0.5080
Fujitsu	fl8wlsb	41	504	3.95	0.54	0.5070
Microsoft	ok8v1	16	131.1	66.5	6.73	0.5280
Microsoft	ok8v2	16	131.1	66.5	5.35	0.5380
Microsoft/CityU	plt8wt1	82	3.04	10.6	1.62	0.5610
UMass	INQ650	215	268	53.8	39	0.5000
UNC	iswqd1	200	40	22	0.005	0.0000
UNC	iswqd2	200	40	22	0.005	0.0000
UWaterloo	uwmt8lw0	5	8.53	32	0.841	0.5720
UWaterloo	uwmt8lw1	5	8.53	32	0.735	0.5580
UWaterloo	uwmt8lw2	5	8.53	32	1.010	0.5650

Table 12: Scale-up factors for CityU/Microsoft and UNC runs for BASE1:BASE10:VLC2.

Group	Measure	Scaleup 1:10	Scaleup 10:100	Scaleup 1:100
CityU/Microsoft	Index_build time	11.6	10.5	122
	Index size	8.23	8.76	72.1
	Query Proc. time	4.32	13.4	57.9
	P@20	1.62	1.29	2.08
UNC	Index_build time			57.1
	Index size			3.67
	Query Proc. time			0.83
	P@20			?

3 Large Web Task

The Large Web Task was by no means as tightly focused as the Small Web Task. A number of different objectives were pursued by the individual participants. They are summarised as follows:

ACSys Investigate the use of link-based measures on the full VLC2 set. Reduce the hardware required for VLC2 processing as much as possible, even to the level of a mid-range laptop. Further study efficiency-effectiveness tradeoffs.

AT&T Test locally distributed IR based on content only.

Fujitsu Labs Comparison of BooleanConjunction+Ranked with Ranked. Efficiency issues. How can VLC2 be indexed using a single index? If multiple concurrent processes on a single processor are used to process queries, what is the optimum degree of parallelism?

Microsoft - Okapi Determine the effect on speed of stop list size, output size limitation, and use of memory vs. use of temporary files.

CityU/Microsoft - Pliers To demonstrate good query processing time and scale-up on a large cluster of machines.

UMass Determine whether UMass conventional retrieval techniques would be effective in the domain of web pages.

UNC Investigate the possibility of having very fast retrieval from a very large information space using a variant of Latent Semantic Indexing.

UWaterloo Fast automatic retrieval on natural language queries. Develop the *cover density ranking* method, using probability based reasoning. Experiment with variations on query length and use of plural/nonplural words in queries.

3.1 Large Web Task: Topics and assessments

ACSys obtained 100,000 "natural language" queries from both Alta Vista [AltaVista Company] and the Electric Monk [Electric Knowledge LLC]. These were censored by a `perl` script to remove possibly offensive queries (or queries which might produce offensive answers)[2] and random selections were made from the remainder until 10,000 queries were selected. These queries were numbered 20001-30000 and distributed to participants.

Participants were required to process all 10000 queries and to submit top 20 rankings to ACSys for judging. After submissions were received, a `perl` script was used to repeat the following until 60 topics had been accepted:

1. randomly select a topic within the 20001 - 30000 range.
2. if the selected topic had fewer than 2 non-stopwords, it was eliminated. The stopword list had 51 entries.
3. If there were at least two non-stopwords, the topic was presented to one of the judges for acceptance or rejection. She was asked to accept a topic if she felt she understood what the person who originally posed the query wanted and if she felt able to judge the relevance or otherwise of documents on that topic.

[2] As became obvious during the ACSys demonstration in the Web Track session at the conference, the list of 110 words to be censored was still missing a few entries!

```
22539 who are the current supreme court justices?
24127 how to make a battery
28771 where can i find the saints and the catholic church?
24111 how to quit smoking?
22905 how to write bibliographies
22719 where can i find information on herbs?
24698 what are the causes of runoff pollution
26776 armstrong louis
21826 where can i find information on the bahamas
24183 how do volcanoes erupt
28150 animal rights
29001 where can i find information about the death penalty?
22674 slobadan milosevic
25597 how do rocks form?
21475 how does a digital camera work?
26981 where can i find information about the civil war
24976 show me a list of vegetarian restaurants in new york city.
22610 thalidomide and multiple sclerosis
25060 old japanese science fictions movies
26274 sinus infection
27375 how do you play chess
29906 where can i find information on the amazon river?
20732 tell me about prozac
26417 how do solar panels work?
24816 hindenburg disaster
28346 find information about american anarchists
26533 why do feet smell?
28850 what are the current ethnic conflicts in azerbaijan?
25358 what are some psychological principles and attitudes for advertising
28273 how to start business
28854 what are the current ethnic conflicts in belarus?
26817 where can i find statistics for education in the united states?
27092 methodist sermons
24790 where can i find information about the politic situation in israel
23274 human genome project
21055 how do i create a web site?
28634 reasons for studying marketing
20784 blood pressure
25233 where can i find information on school violence?
21247 where can i find information on russia?
28677 where can i watch tv on the internet?
21185 where can i find the best jokes?
25663 how are hospitals prepared for y2k?
28798 where can i find information about teenage alcohol abuse in the uk
28846 teen alcohol abuse statistics for the uk
27358 egyptian history
```

Figure 4: A sample of the judged queries used in the Large Web task. Note the two very similar `"teen alcohol"` queries at the bottom of the list. Note also the retention of probable query errors: `"slobadan"`, `"science fictions"`, and `"politic situation"`.

Something less than 100 selections were eliminated and 718 were rejected by the judge. Of the 60 accepted, it transpired that our judge had accidentally accepted one, `27188 where can i find black escorts?` which she really wanted to reject and she also accepted two topics likely to have the same answers: `24111 how to quit smoking? / 23728 how do i quit smoking?` Topics 27188 and 23728 were rejected.

Sample accepted queries are shown in Figure 4. Note that two similar queries related to "teen alcohol" were both accepted.

The number judged dropped by a further one when one of the judges was unavailable for a few days after all other topics were finished. The resulting 57 topics were pruned to 50 by arbitrarily eliminating all the topics for which there were fewer than 5 relevant documents.

The pooled documents for each topic were presented to the assessors in order of increasing document length using the RAT (Relevance Assessment Tool) used in previous VLC track experiments. This time however, a text-only web browser [Lynx] was used to display documents in a way which rendered references and tables in a reasonable way (minus images).

The six assessors were all University graduates from specialties other than Computer Science or Librarianship. Three of them had served as VLC track judges in previous years.

3.2 Large Web Task: Efficiency-effectiveness results

The tradeoffs between efficiency and effectiveness are actually tradeoffs between five dimensions:

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

For each run submitted against the full 18.5 million document collection, Figures 5 and 6 show a 5-axis Kiviat diagram summarising performance on each of these dimensions. 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 corrsponds to minimum score.

To illustrate the scaling process, the smallest index size was achieved by Fujitsu at 3.9 gigabytes. 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.

Use of linear scaling in the Kiviat diagrams tends to exaggerate the differences between runs, whereas log scaling would have tended to homogenize them. The shape of the diagram indicates the degree to which that run achieved good performance (relative to the group) on one (or a couple of) dimensions at the expense of the others, or alternatively achieved a good balance between them. Good balance is indicated by a filled-out shape, best illustrated by the hypothetical "uniformly best" system shown at the top left of Figure 5.

The Kiviat diagrams shown in Figure 5 are considerably distorted by the inclusion of the UNC runs which achieved enormous query processing speed but very low precision. The diagrams in the results section of the TREC-8 proceedings are quite different because they do not include the UNC runs (which were submitted after the deadline.)

3.3 Large Web Task: Scalability results

Two groups submitted scalability runs in which the VLC performance figures were compared with those of the BASE1 and BASE10 uniform samples. The scaleup factors for these runs are presented in Table 12.

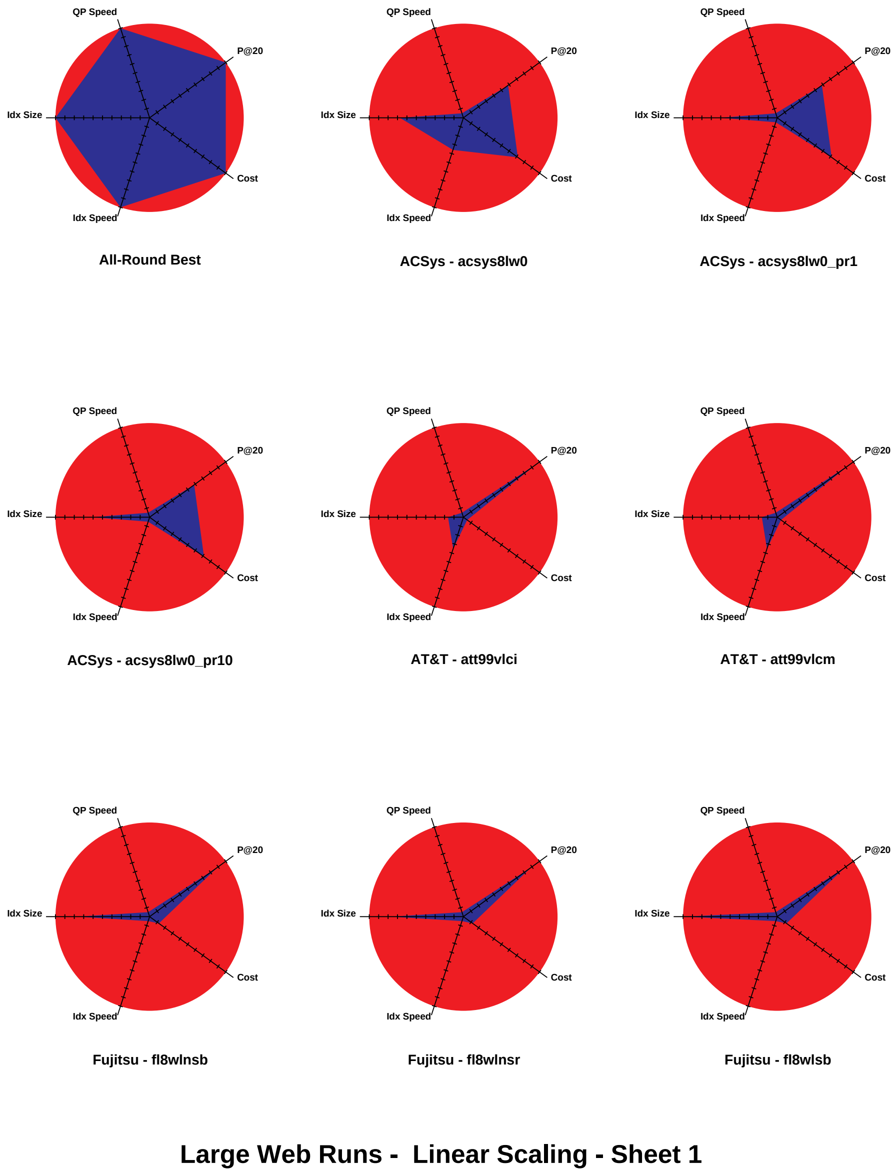

Figure 5: Composite results for all runs submitted in the Large Web Task. Note that the UNC runs were submitted after the deadline and consequently included a very high percentage of unjudged documents. Accordingly, their precision result is very low. However, their query processing was two orders of magnitude faster than the next best 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. (Continued in Figure 6.)

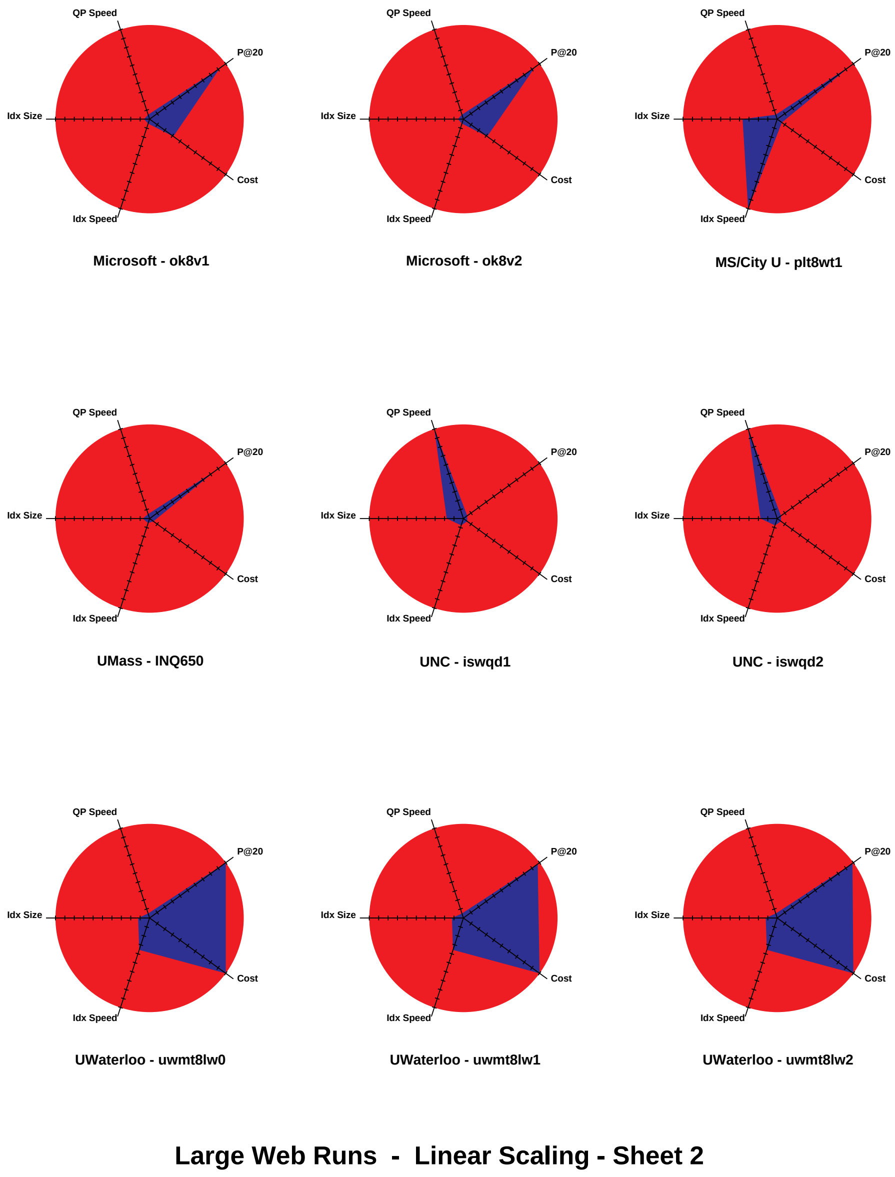

Figure 6: Continuation of Figure 5.

3.4 Large Web Task: Exploiting links

ACSys (in the person of Nick Craswell) computed PageRank scores for all the documents in the VLC2. Results are reported in the ACSys paper in these proceedings.

In essence, computation of PageRanks took much longer than indexing but use of PageRanks increased query processing time only slightly (by an average of 0.15 sec per query, less than 5%.) However, the benefit in terms of query processing effectiveness was found to be negligible.

3.5 Large Web Task: Hardware resources

Several groups were successful in reducing the scale of hardware required to process the full VLC2 collection, compared to what they used in TREC-7 VLC track. UWaterloo reduced their machinery from four PCs to two. ACSys used one PC instead of eight DEC Alphas and demonstrated query processing over the full collection on a Dell laptop using only the internal disk drives.

Fujitsu (who did not participate in TREC-7 VLC) demonstrated that, by eliminating non-English documents and HTML tags, the whole of the VLC2 could be represented in a single index of only 3.9 gigabytes.

3.6 Large Web Task: Other issues

The various other questions addressed by participants are covered in their own papers.

Acknowledgements

We are very much indebted to Brewster Kahle of the Internet Archive for making available the spidered data from which the VLC2 and WT2g collections are derived and to Alta Vista (Monika Henzinger and Michael Moricz) and the Electric Monk (Edwin Cooper) for providing large samples of queries from their logs. Thanks also to John O'Callaghan and Darrell Williamson (successive CEOs of ACSys) and Peter Langford (ACSys Centre Manager) for supporting the track.

Finally, thanks are due to Sonya Welykyj, Penny Craswell, Clare Dyson, Zoe McKenzie, Greg Evans and Nick Clarke for their work in assessing Large Web submissions.

Bibliography

ACSys. TREC Web Tracks home page. `http://pastime.anu.edu.au/TAR/WT`.

AltaVista Company. Alta Vista web page. `http://www.altavista.com/`.

Electric Knowledge LLC. Electric Monk home page. `http://electricmonk.com/`.

Hawking, D., Craswell, N., and Thistlewaite, P. 1998. Overview of TREC-7 Very Large Collection Track. In E. M. Voorhees and D. K. Harman Eds., *Proceedings of TREC-7* (Gaithersburg MD, November 1998). `http://trec.nist.gov/pubs/trec7/t7\_proceedings.html`.

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

NIST. Guide to Z39.50/PRISE 2.0. `http://www.itl.nist.gov/div894/894.02/works/papers/zp2/zp2.html`.