Overview of the TREC-9 Web Track

David Hawking[*]
CSIRO Mathematical and Information Sciences,
Canberra, Australia
David.Hawking@cmis.csiro.au

September 4, 2001

Abstract

TREC-9 marked a broadening of the range of of search task types represented in the Web track and a serious attempt to determine whether hyperlinks could be used to improve retrieval effectiveness on a topic-relevance ad hoc retrieval task. The Large Web Task compared the ability of systems to locate online service pages within the 18.5 million page VLC2 collection. In this case the question is not whether the page is relevant to the topic, but whether it provides direct access to the desired service. In contrast, the Main Web Task compared link-based and non-link methods on a task involving topic relevance queries and a 1.69 million page corpus (WT10g) which was carefully engineered to ensure a high density of inter-server links and (relative) ease of processing. The Main Web task topics were in TREC Ad Hoc form but were reverse engineered from query logs. Ternary relevance judgments were obtained and, in addition, assessors were asked to identify "best" documents for each topic. As in TREC-8, no significant benefit associated with the use of link information in a topic-relevance retrieval task was demonstrated by any of the participating groups, whether or not additional weight was given to highly relevant documents.

1 Introduction

The TREC-9 Web Track activities centred on two tasks: the Main Task and the Large Task. The latter made use of the 100 gigabyte, 18.5 million webpage VLC2 collection described in the 1998 VLC Track overview [Hawking et al. 1998]. The former worked with a 10 gigabyte, 1.69 million document subset of the VLC2, distributed on five CD-ROMs as the WT10g collection. [Bailey et al. 2001].

A final Web Track activity was the invited talk to a TREC-9 plenary session by Dr Andrei Broder, Chief Scientist at the Alta Vista search engine company. Slides from an updated version of that talk, presented at the 2001 Search Engines Meeting, by Andrei's colleague Bob Travis, are available online [Travis and Broder 2001].

2 Main Web Task

None of the participants in the TREC-8 Small Web Task, using a two gigabyte corpus (WT2g), managed to demonstrate any benefit whatever from using hyperlink methods in that particular retrieval task. Given that most commercial Web search engines exploit hyperlinks to apparently great effect, this result may seem surprising.

Accordingly, a new task was devised for TREC-9 which removed possible impediments to good link-based performance which were perceived to be inherent in last year's task.

[*] The author wishes 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.

2.0.1 Main Web: Task Summary

Corpus: A new test corpus was deliberately constructed in such a way as to dramatically increase the density of inter-server hyperlinks. Full details of the construction of the new corpus (known as WT10g) are documented elsewhere [Bailey et al. 2001]. WT10g was defined and initially distributed by the ACSys Cooperative Research Centre. Following the demise of ACSys in September 2000, WT10g has been distributed by CSIRO[1], which was one of the ACSys partners.

Note that, although WT10g is a subset of VLC2, documents in WT10g are assigned different document numbers to enable easy extraction of the document. However, they include the original document numbers within `<DOCOLDNO> ... </DOCOLDNO>` tags.

Topics: Test topics were reverse engineered from queries selected from real Web search engine[2] logs, rather than being generated with respect to the Ad Hoc corpus. The topics were presented in traditional Ad Hoc form with title, description and narrative fields. The title field contained the unedited query from the original log. The description and narrative fields were a statement of a particular interpretation of the query to be used in judging. For example, the polysemous query `cats` might have been interpreted as `Who wrote and who acted in the musical "Cats"?`. Documents relating to other meanings such as bulldozers and domestic pets would be judged as irrelevant.

Misspellings were a feature of several queries chosen as topics. An example was the single word query `angioplast7`. Not surprisingly, there were no occurrences of this word in the collection. Successful processing of the title-only version of this topic thus required either spelling-correction, approximate matching, or n-gram methods.

Results required: Participants were required to return a top 1000 (or fewer) list of documents for each topic, ranked in order of decreasing estimated relevance to the topic.

Judments: Ternary (irrelevant/relevant/highly relevant) rather than binary judging was adopted. Judges were also asked to identify *best* documents from among the highly relevants. Two additional judges were later asked to examine the relevant and highly relevant documents for each topic and to pick what they considered to be the best one (or possibly more than one.)

Focus: The principal focus of the Main Web Task was to re-attempt last year's question, in more favourable circumstances:

- Can link information in Web data be used to obtain more effective search rankings on a topic-relevance Ad Hoc retrieval task than can be obtained using page content alone?

2.1 Types of Run

A strong distinction must be drawn between automatic, title-only (*short*) runs and the rest (*notshort*). In real Web search, search engines only have access to the query recorded in the title field. The underlying information need is known only to the searcher and not to the search engine. Thus, only the short runs are representative of real Web search.

Despite this, notshort runs were encouraged because they add value to the test collection by increasing the number of known relevant documents. They also give some idea of what level of performance may be possible on each task and allow groups to continue work on longer queries.

The notshort category includes: interactive manual, blind manual, and automatic runs which used any part of the topics other than the title.

2.2 Judging pools

The number of runs judged was 59, giving a maximum pool size of 5900 (per topic). The mean actual pool size was 1401, 23.8% of the maximum, while the mean number of relevant documents over the 50 topics was 52.34, or 3.7% of the number of documents judged.

[1] Commonwealth Scientific and Industrial Research Organisation, an Australian government agency. (`mailto://test_collections@act.cmis.csiro.au`)

[2] eXcite

53.3% of the relevant documents were returned by both automatic and manual runs. 10.4% of the relevants were found by manual runs only, and another 15.3% were found by notshort runs only. Thus, if all runs other than the short ones were excluded, one quarter of the relevant documents would have been lost. Automatic runs (only) contributed 68.5% of the pools, manual (only) 15.6% of the pools, and so 15.8% of the pools were contributed by runs of both types.

Twenty-three groups submitted at least one run to the main web track, but only 21 submitted in time for runs to be judged. All of the 21 groups retrieved at least eight relevant documents in the top 100 across all 50 topics that no other group retrieved (*unique relevants*). The group that found the most unique relevants was Illinois Institute of Technology, with 212 over all topics. 162 of the 212 came from their manual run. The group with the next highest number of unique relevants was Hummingbird with 57 (over all topics). Justsystems and CUNY each had 55, with no one else having more than 50.

A total of 105 runs were submitted. Of these, 78 were content-only and 27 were content-link. There were 12 manual runs, 50 automatic short runs, and 53 automatic notshort runs.

2.2.1 Pool completeness - 1

NIST recently conducted experiments to determine what the effect of pooling fewer documents from each run would have been. Two cases were compared with official pools were based on the top 100 documents: top 50, and top 75.

Here are the results of that investigation:

Pool	Average pool size	Average number relevant
Top 100	1401.44	52.34(100%)
Top 75	1077.12	46.52 (88.88%)
Top 50	743.20	39.64 (75.74%)

NIST have decided on this basis to continue basing topic relevance pools on the top 100 documents. The drop in number relevant for 75 is not serious, but the pool size isn't reduced enough to be worthwhile (and one topic that had only 3 relevant would have lost 1). The drop in number of relevants for the top 50 case was considered too severe.

Unfortunately, it is not possible to reliably quantify what the effect of increasing the number of documents in the pools would have been. The following table shows how the probability that a pooled document will be judged relevant depends upon the rank at which it was retrieved.

Rank range	docs judged relevant
1-50	5.33%
51-75	2.06%
76-100	1.70%

Based on the very crude approximation to this data shown in Figure 1, we estimate the probability of documents in the 101-200 range being judged relevant if they had been included in the pool as 0.97%, suggesting that doubling the depth of judging might have increased the number of relevant documents per topic by an average of about 13 (about 25%).

It therefore seems almost certain that there are unjudged relevants within the collection. However, averaged over 50 topics, these are unlikely to affect relative system rankings.

2.2.2 Pool completeness - 2

Another way to look at pool completeness is to see how much judged runs' evaluations differ when using qrels with and without that group's unique relevants. NIST ran this computation for the TREC-9 web track (using both levels of relevant as "relevant") and mean average precision. The results are encouraging.

The largest percentage difference in mean average precision (MAP) is 10.3, but that run had poor effectiveness (MAP of .0174 using original qrels) and the absolute difference was only .0018. The next highest percentage difference was 6.1, again for a poor run. The third highest percentage difference was

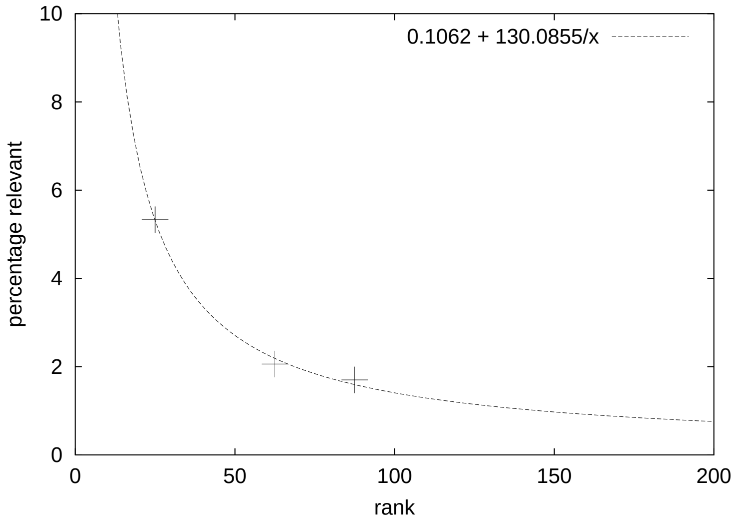

Figure 1: Decline in probability of document being judged relevant with increasing rank at which it was retrieved, shown with the $y = b + c/x$ line of best fit.

5.2 for the IIT manual run, the run which contributed the most unique relevants. For runs with a MAP of at least .1, the percentage difference was almost always less than 2%, except for some TNO runs that actually improved by around 3% when evaluated without their own unique relevants! When the process was repeated using P@10 instead of MAP, 38 of the 59 runs showed no change whatsoever. The biggest percentage difference was 7.72 for the IIT manual run.

2.2.3 Pool completeness - 3

How much of a problem is the presence of unjudged relevant documents in a test collection? Zobel [Zobel 1998] conducted various tests to try to determine how incomplete TREC collections relevance judgments are. He found that there were unjudged relevant, that the number of unjudged relevant was highly skewed by topic (the more relevant in early ranks, the more relevant there continues to be), and that the quality of the pools (diversity of systems contributing to the pools and depth in the system's ranking) did affect the quality of the resulting collection. But he also found that the TREC collections he looked at were quite acceptable for comparing retrieval systems – the errors he observed due to incompleteness were smaller than the differences occasioned by using different relevance assessors.

2.3 Properties of WT10g

In summary, WT10g is considerably larger than earlier TREC ad hoc collections and WT2g. However, ease of processing was improved by elimination of many of the binary and Non-English pages normally found in Web crawls. Most importantly, WT10g includes a very much higher density of inter-server hyperlinks than did WT2g. Readers are referred to [Bailey et al. 2001] for full WT10g collection properties.

Table 1 compares the densities of known relevant documents for the TREC-9 Main Web topics with that of other recent TREC collections. Naturally, there may be considerable variation from one topic to another.

2.3.1 Connectivity data

Nick Craswell's software for extracting hyper-link connectivity information from collections was run over WT10g and the resulting connectivity matrix was distributed with the collection on CD-ROM.

Table 1: The density of known relevant documents (across 50 topics) in WT10g for the TREC-9 topics compared to that in earlier tasks.

Judgments	Collection	Density of relevant docs
T7	VLC2	6482/18.5M = 0.03 %
T8	WT2g	2279/247491 = 0.92 %
T8	Ad Hoc	4728/528155 = 0.90 %
T9	WT10g	2371/1.69M = 0.14 %

2.4 Summary of participation

Tables 2 and 3 list the official runs submitted in the Short and Notshort categories. Runs which made use of link information are marked with the word LINK. Table 4 summarises the methods used by the Main Web participants.

2.4.1 Link v. Content

JustSystem Some pairs of runs showed an advantage arising from the use of anchor text on some measures. The biggest advantage was a few percent superiority for the `jscbt9wll2` run against the `jscbt9wcl1` baseline on all three measures. However, JustSystem conclude from a large set of unofficial as well as official runs that the benefit is small and inconsistent.

U Waterloo A minute gain on average precision was reported from the use of an unspecified link method for the T+D runs. In all other cases, use of links caused harm.

AT&T Use of relevance feedback `att0010gbt` harmed both early precision and average precision. The runs `att0010glf` and `att0010glv`, which used anchortext, performed worse on these measures than all the content-only runs except the relevance feedback run. Interestingly, there is no drop in performance for `att0010gbt` when runs are compared using DCG[100]. Upweighting query words in title fields (`att0010gbe`) was beneficial on all measures.

Other Runs Inspection of Tables 2 and 3 reveals no other indication of benefit achieved from use of link-based methods.

It is clear that this data and participant reports confirm last year's observation that no consistent benefit was gained from the use of links in this topic-relevance ad hoc retrieval task.

This seems to be also true, even when highly relevant documents are valued very highly.

2.4.2 Resilience to Query Misspelling

The effect of misspelled query words should be most noticeable in the short category. The five top-performing groups on that category used the following approaches:

JustSystem If insufficient results were returned in response to the original query, the query was automatically expanded to include spelling variations.

Hummingbird If an original query term occurred in fewer than ten documents, then Soundex-matching approximate match terms from the collection, with up to 2 edit errors were added. If the number of documents affected was still less than ten, then non-Soundex approximate matches were added. If necessary, trailing letters were dropped until the ten-document criterion was satisfied.

U Waterloo They found that 4-gram indexing worked better than words partly because of misspelling resilience but also because of conflation of morphological variants.

Table 2: All official runs submitted in the short (automatic, title-only) category, presented by group. Groups are ranked in order of decreasing average precision of their best run. The DCG[100] figures represent the discounted cumulative gain (described in the text) when a highly relevant document is considered to be worth 100 times as much as a relevant one. Natural logarithms were used ($b = e$) and the cutoff was at rank 100. The number of groups was 19 and the number of runs submitted was 40.

Group	Run tag	Ave. prec.	P@10	DCG[100]	Type
JustSystem	jscbt9wcs1	0.2011	0.238	107.024	
	jscbt9wls1	0.2000	0.252	110.027	LINK
	jscbt9wls2	0.1838	0.224	96.319	LINK
Hummingbird	hum9te	0.1970	0.254	118.725	
U Waterloo	uwmt9w10g0	0.1654	0.238	95.620	
	uwmt9w10g2	0.1631	0.236	95.356	LINK
	uwmt9w10g4	0.1812	0.240	94.366	
	uwmt9w10g5	0.1794	0.240	93.341	LINK
Twenty-One	tnout9t2	0.1801	0.214	107.745	
	tnout9t2lc10	0.1630	0.214	106.238	LINK
	tnout9t2lc50	0.1337	0.198	102.628	LINK
	tnout9t2lk50	0.0488	0.032	29.912	LINK
Ricoh	ric9tpx	0.1787	0.276	118.981	
U Neuchatel	NEtm	0.1754	0.212	104.484	
	NENRtm	0.1743	0.208	103.373	
	NENRtmLpas	0.1736	0.208	103.214	LINK
Queens CUNY	pir0Wt1	0.1750	0.218	99.975	
IIT/AAT/NCR	iit00t	0.1627	0.250	109.718	
ATT	att0010gb	0.1341	0.200	89.104	
	att0010gbe	0.1464	0.226	101.331	
	att0010gbl	0.1380	0.204	85.803	
	att0010gbt	0.1182	0.172	89.870	
	att0010glf	0.1250	0.182	89.579	LINK
	att0010glv	0.1288	0.190	90.983	LINK
Fujitsu Labs	Flab9atN	0.1360	0.194	102.861	
JHU/APL	apl9lt	0.1062	0.116	64.291	LINK
	apl9t	0.1272	0.134	77.119	
SABIR Research	Sab9web1	0.1265	0.182	98.120	
IRIT	Mer9Wt0	0.0996	0.128	56.285	
	MEr9Wt1	0.0114	0.070	14.546	
Seoul National U	Scai9web3	0.0915	0.154	83.513	
U Padova	PuShortAuth	0.0591	0.136	72.266	LINK
	PuShortBase	0.0654	0.142	74.769	
	PuShortWAuth	0.0637	0.138	73.591	LINK
UNC	isnnwt	0.0225	0.058	28.626	
	iswt	0.0240	0.076	38.807	
Pam Wood	UCCS1	0.0181	0.052	15.037	
	UCCS2	0.0169	0.052	16.792	
CWI	cwi0000	0.0176	0.066	13.609	
	cwi0010	0.0125	0.024	9.428	

Table 3: Measures as for Table 2. All official runs submitted in the notshort (manual, interactive, or non-title-only automatic) category, presented by group. Groups are ranked in order of decreasing average precision of their best run. The number of groups was 22 and the number of runs submitted was 65.

Group	Run tag	Ave. prec.	P@10	DCG[100]	Type
IIT/AAT/NCR	iit00td	0.2293	0.350	143.533	TD
	iit00tde	0.2217	0.346	143.306	TD
	iit00m	0.3519	0.518	172.199	M
JustSystem	jscbt9wcl1	0.2687	0.342	135.893	TDN
	jscbt9wll1	0.2659	0.344	132.645	TDN/LINK
	jscbt9wll2	0.2801	0.358	144.059	TDN/LINK
Ricoh	ric9dpn	0.2616	0.338	151.226	TD
	ric9dpx	0.2267	0.322	136.544	TD
	ric9dsx	0.2201	0.324	137.191	TD
	ric9dpxL	0.2257	0.316	135.891	M
U Neuchatel	NEnm	0.2499	0.342	142.946	TDN
	NEnmLpas	0.2488	0.340	142.417	TDN/LINK
	NEnmLsa	0.2185	0.332	136.446	TDN/LINK
ANU/CSIRO	acsys9mw0	0.2486	0.384	144.506	M
Hummingbird	hum9td4	0.2115	0.308	127.553	TD
	hum9tde	0.2217	0.294	121.085	TD
	hum9tdn	0.2335	0.352	139.349	TDN
Queens CUNY	pir0Wtd2	0.2164	0.302	122.122	TD
	pir0Wttd	0.2097	0.318	92.404	TD
	pir0WTTD	0.1418	0.180	92.404	TD/LINK
	pir0Watd	0.2209	0.298	130.067	TDN
Twenty-One	tnout9f1	0.2178	0.290	132.224	TDN
NeurOK	NRKlm	0.2064	0.282	126.202	TDN
	NRKprf20	0.2173	0.326	131.539	TDN
	NRKse10	0.1960	0.272	117.767	TDN
	NRKse20	0.1642	0.234	108.058	TDN
SABIR Research	Sab9web2	0.2122	0.340	134.015	TDN
	Sab9web3	0.2159	0.346	135.596	TDN
	Sab9web4	0.2091	0.342	134.333	TDN
	Sab9web5	0.2018	0.314	126.863	TDN/LINK
JHU/APL	apl9td	0.1917	0.340	120.747	TD
	apl9all	0.1948	0.314	125.502	TDN
	apl9ltdn	0.1494	0.232	98.921	TDN/LINK
	apl9tdn	0.1785	0.286	115.197	TDN
Fujitsu Labs	Flab9atd2N	0.1877	0.302	132.741	TD
	Flab9atdN	0.1816	0.298	139.320	TD
	Flab9atdnN	0.1923	0.316	139.320	TDN
SUNY Buffalo	xvsmmain	0.1521	0.214	110.224	TD
	xvsmtitle	0.1278	0.184	92.268	TN
	xvsmtdn	0.1694	0.238	113.566	TDN
	xvsmman	0.1785	0.260	119.496	M
Dublin City U	dcu00ca	0.1519	0.278	117.162	M
	dcu00la	0.1450	0.274	111.980	M/LINK
	dcu00lb	0.1324	0.244	102.915	M/LINK
	dcu00lc	0.1387	0.258	107.824	M/LINK
U Waterloo	uwmt9w10g1	0.1331	0.260	87.252	TD
	uwmt9w10g3	0.1336	0.262	87.860	TD/LINK
Seoul National U	Scai9Web1	0.0941	0.152	84.387	TD
	Scai9Web2	0.0934	0.138	82.733	TD
	Scai9Web4	0.0946	0.146	84.214	TD
RMIT/CSIRO	rmitNFGweb	0.0707	0.088	56.950	M
	rmitNFLweb	0.0702	0.090	54.944	M
	rmitWFGweb	0.0040	0.010	11.663	M
	rmitWFLweb	0.0341	0.044	27.193	M
U Padova	PuLongAuth	0.0648	0.172	81.156	TD/LINK
	PuLongBase	0.0666	0.180	83.536	TD
	PuLongWauth	0.0660	0.178	83.144	TD/LINK
UNC	iswtd	0.0325	0.110	42.449	TD
	iswtdn	0.0412	0.084	32.362	TDN
CWI	CWI0001	0.0174	0.054	14.872	TD
	CWI0002	0.0122	0.038	8.110	TDN
IRIT	Mer9WtdMr	0.0154	0.090	15.548	TD
	Mer9Wtnd	0.0140	0.092	14.758	TDN
Pam Wood	UCCS3	0.0000	0.000	0.000	D
	UCCS4	0.0000	0.000	0.000	D

Twenty-One Fuzzy matching?

Ricoh No reported correction.

2.4.3 Value of Query Expansion

Again considering only the short runs, the five top-performing groups reported the following:

JustSystem Found consistent improvement in average precision using both reference database feedback and pseudo-relevance feedback on WT10g. The combination achieved a gain of 16-17%.

Hummingbird An expansion method similar to Rocchio produced very small gain when evaluation was based on all relevant documents and caused harm when only highly relevants were considered.

U Waterloo No feedback employed.

Twenty-One Training with WT2g revealed that blind feedback caused harm due to large numbers of typographical errors in documents.

Ricoh In the official runs, query expansion caused harm. However, subsequent unofficial runs with a modified expansion method showed an improvement of 8% in average precision.

2.4.4 Short vs. Notshort

Table 5 compares the performance of the group of short runs versus the group of automatic notshort runs. As can be seen, there are substantial differences in favour of the notshort group. Comparing the medians for the groups, notshort is 49% better than short on P@10 and 44% better on average precision.

Best performance overall was achieved by the manual run `iit00m` from IIT/AAT/NCR. Its P@10 score was 88% better than the best P@10 score for a short run (`ric9tpx`, submitted by Ricoh). Its average precision score was 75% better than the best for a short run (`jscbt9wcs1`, submitted by JustSystem).

Note that the best possible P@10 score was 0.878, due to topics with less than 10 known relevant documents.

2.4.5 Evaluation by Highly Relevant Documents

Voorhees [Voorhees 2001] found that WT10g system rankings based on Highly Relevant judgments were non-trivially different from those based on Relevant and Highly Relevant combined. Because the numbers of Highly Relevant documents are relatively small she recommended the use of the *Discounted Cumulative Gain* (DCG) method proposed by Järvelin and Kekäläinen [Järvelin and Kekäläinen 2000] to combine information from both categories of relevance, but with higher weight to the Highly Relevants.

Tables 2 and 3 report DCG scores for the Main Task runs with a heavy bias toward Highly Relevant.

2.4.6 Evaluation by Best Documents

Voorhees [Voorhees 2001] found little agreement between judges about which pages were the best resources on a topic. She found that best page judgments did not lead to stable measures.

However, the best page judgments give the opportunity to look at whether the best resources on a topic tend to be: a) site homepages, b) close to the root of a directory tree, c) documents with higher than average in-link count. The following analysis uses the union of the sets of bestpage judgments for each of the three judges.

Site homepages: I examined each of the 130 documents identified by one or more of three judges as the best for a topic. I classified only one of them as a site home page (a small site published by the Rainbird Company about the Rose Parade). There were also three Yahoo! directory pages but most of the best documents were individual pages which presented detailed information on the topic.

Depth in directory hierarchy: On average, the best pages were 2.61 levels deep within the directory hierarchy, compared with an average of 2.96 levels for all pages in the collection.

Table 4: Style of link exploitation methods used by groups participating in the Main Web Task. In many cases, the methods actually employed represent modifications of the basic method listed.

Group	Methods
ATT	Anchor Text Propagation
Dublin City U	Inlink Count
Dublin City U	Spreading Activation
Dublin City U	HITS/Co-citation
JHU/APL	Inlink Count
JustSystem	Anchor Text Propagation
Queens College CUNY	Inlink Count
SABIR	Not Stated
Twenty-One	HITS
Twenty-One	Co-citation
U Neuchatel	Spreading Activation
U Neuchatel	Probabilistic Argumentation
U Neuchatel	HITS
U Neuchatel	PageRank
UPadova	HITS
UPadova	HITS/Similarity
U Waterloo	Not Stated

Table 5: Comparative performance of short v. notshort runs (excluding manual runs).

Measure	Short	Notshort
P@10 (best)	0.276	0.358
P@10 (median)	0.198	0.296
Ave Prec (best)	0.2011	0.2801
AvePrec (median)	0.1341	0.1936

Directory depth	1	2	3	4	5	6	7	8
No. of Bests	32	32	38	14	13	-	-	1

In-link counts: Based on the `in_links.gz` file distributed with WT10g, there are 8,062,918 inlinks across the 1,692,096 documents in the collection, giving an average number of inlinks per page of 4.77. Nick Craswell has computed inlink counts for each of the judging categories.

Category	All docs	Docs Judged Irrelevant	Docs Judged Relevant	Docs Judged Highly Rel.	Bests
Mean Inlinks/page	4.77	8.48	5.26	4.80	4.67
Median Inlinks/page	1	1	1	1	2

The inlink distribution for the collection is very heavily skewed, with a high peak at one. The fact that documents judged to be irrelevant show a much higher inlink density than the average for the collection, probably results from the unsuccessful use of link-based methods by participants. It is quite interesting that relevant and highly relevant documents are not distinguished from randomly chosen documents by their inlink count. Indeed the fact that the median for both groups is 1 indicates that more than half of them have 1 or fewer inlinks.

Superficially, it may seem that there is an exploitable difference between the inlink counts for the best pages and for pages in general. The raw average inlink count for best pages is 9.78, however that figure is grossly distorted by an extreme outlier with 337 incoming links. That single page accounts for 53% of all incoming links to best pages. Excluding it reduces the average to 4.67, the figure reported in the table. In fact, the median score for the bests is barely above one.

It appears that best pages are not distinguishable in any useful way from pages in general or from less relevant pages on any of the three attributes considered.

2.5 Main Web Task discussion and conclusions

Evidence is already available [Bailey et al. 2001] that WT10g is in fact large enough and contains sufficient links to demonstrate a dramatic advantage to a link based method on a homepage finding task. It seems reasonable to conclude that link methods can be beneficial in some forms of retrieval task, even over a small test collection like WT10g.

However, no such advantages have been found for topic relevance tasks. Even though consideration of multiple degrees of relevance does change relative system rankings, it still does not allow demonstration of any worthwhile advantage to link methods. Indeed, even the best resources for a topic appear not to be usefully distinguishable by frequency of incoming links.

3 Appropriate Web Evaluation Methodology

Following robust and mutually beneficial debate at the Infonortics Search Engines 2000 meeting[3], the Web Track organisers became convinced of the need to significantly extend TREC Ad Hoc evaluation methodology to accommodate different types of retrieval task and to use appropriate judging instructions and measures for each type.

TREC Ad Hoc retrieval exercises (including previous VLC and Web tasks) have concentrated on topic relevance tasks in which a searcher is assumed to be looking for a range of documents which are relevant to a particular research topic. Such retrieval tasks are definitely represented in Web search engine query streams but form only a small proportion of the total.

Following the Search Engines 2000 meeting, CSIRO/ANU have proposed a taxonomy in which search tasks are classified at the top level by how many results the searcher expects:

[3] `http://www.infonortics.com/` The debate particularly involved David Hawking and Chris Buckley representing TREC evaluation and (among others) Larry Page of Google, Eric Brewer of Inktomi and Andrei Broder of Alta Vista, representing Search Engine companies. David Evans was the moderator.

- a part document (eg. Q&A),
- a single document (eg. known-item and homepage search),
- a selection of documents (This class may include topic relevance and online-service location tasks), and
- all documents matching a criterion (eg. metadata search, such as all documents authored by a particular person).

CSIRO/ANU have conducted evaluations of public search engines using on-line service location (selection of pages) and homepage finding (single page) [Hawking et al. 2001; Hawking et al. 2001; Craswell et al. 2001] as well as topic relevance tasks. On-line service evaluation was also used in the TREC-9 Large Web task.

In his TREC-9 talk, Andrei Broder presented an alternative taxonomy of Web search, dividing searches into:

- Informational,
- Transactional, and
- Navigational

classes, each of which require different evaluation methods.

The TREC-2001 Web Track includes a homepage finding task.

4 Large Web Task

In the past, participation in the VLC track and the Large Web track was limited by the difficulties of processing 100 gigabytes of data. This year however, the number of participants declined, due to the fact that 18.5 million pages (100 gigabytes) no longer constitutes an interesting challenge to those seriously pursuing scalable, large scale retrieval. At the time of the TREC-9, major Web search engines were indexing around 30 times as many pages as are contained in the VLC2 collection.

A number of different objectives were pursued by the individual participants. They are summarised as follows:

ACSys Comparing anchor text and PageRank resorting methods with Okapi BM25.

AT&T Testing the new retrieval system Tivra. Comparing anchor text with content only.

Fujitsu Labs Engineering to achieve a good balance between speed, effectiveness and cost.

Hummingbird Evaluating an experimental version of Fulcrum SearchServer on large data. Testing approximate search.

U Waterloo Repeat of TREC-8 runs.

4.1 Large Web Task: Topics and assessments

The 10,000 "natural language" queries from the TREC-8 Large Web task were re-used. They were obtained by random selection from combined large Alta Vista and Electric Monk search logs and were numbered 20001-30000.

Participants were required to process all 10000 queries and to submit top 10 rankings to ACSys for judging. After submissions were received, the track coordinator (David Hawking, who was not an official participant in the TREC-9 Large task) selected 106 of the topics which seemed to have been intended to locate some form of online service. Four of these were used as practice by the judges. Sample accepted queries are shown in Figure 2.

The pooled documents for each topic were presented to the assessors in order of increasing document length using the RAT (Relevance Assessment Tool [Hawking et al. 2001]) 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 four assessors were all University graduates from specialties other than Computer Science or Librarianship. Two of them had served as VLC track judges in previous years.

During judging zero "good" documents were found for 18 queries, which are therefore excluded from the following analysis.

The number of good documents found per query ranged from 1 to 72, with a mean of 24.1. The number of queries for which fewer than five good documents were found was 14 and 21 had fewer than 10. A total of 6911 documents were judged.

4.2 Runs

All runs judged in the Large Task are listed in Table 6. The P@10 results are also shown graphically in Figure 3. Figure 4 shows the tradeoffs between cost, speed, space and effectiveness for the runs for which detailed information was provided.

The runs in Table 6 which are labelled with an asterisk were "submitted" by Nick Craswell who was a PhD student in ACSys (to June 2000), then an intern at Microsoft Research, Cambridge (June-September 2000) and subsequently an employee of CSIRO. The `acsys9*` runs constituted an experiment in use of hyperlink methods and were in fact judged blindly but cannot be fairly compared with those of other participants.

Nick used the TREC-8 Microsoft Research run as a baseline. This was possible because TREC-8 Large Task participants had submitted results for all 10,000 queries, not just the 50 or so which were judged in TREC-8. The run `acsys9pr` was a PADRE run in which the top 1000 documents were reranked on the basis of PageRank scores computed for the VLC2 collection. The other two ACSys runs made use of link anchor text. All links whose anchor text matched the query were included and documents were scored on the basis of a count of incoming matching links. Run `acsys9lnkA` included all matching links whereas `acsys9lnkE` excluded within-site links.

4.2.1 Large Web Notes

Hardware/OS AT&T, Fujitsu and Hummingbird all used single low-cost PC systems. AT&T and Fujitsu used Linux and Hummingbird used Windows NT.

Cost Fujitsu 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.

Comparison with TREC-8 The Microsoft run evaluated in both TREC-8 and TREC-9 achieved a P@10 of around 0.56 in TREC-8 and a slightly lower figure of 0.52 in TREC-9. By contrast, the University of Waterloo TREC-8 runs achieved 0.58-0.59 but identical runs in TREC-9 came in at only 0.43-0.45. The differences between Waterloo and Microsoft on TREC-9 have not been subjected to statistical significance testing but if the differences were significant it would be interesting to investigate whether the algorithms used by Waterloo worked better on a topic relevance rather than an online-service finding task.

Link methods AT&T were surprised to find no benefit from use of anchor text in this task. ACSys also found no benefit from anchor text or PageRank reranking relative to the Microsoft baseline, but the scoring methods used in the anchor text case were not very sophisticated.

5 Conclusions

No conclusive or consistent benefit from the use of link information was demonstrated on the Main Task, despite a larger (10 gB) dataset, a relatively high density of inter-server links and the use of assessments

```
25209 where can i find love songs?
25363 where can i find cd rom drivers
25418 any packaging business for sale?
25538 where can i find some frank frazetta wallpaper?
25744 mp3
25819 where can i shop for toys?
25861 where can i learn dutch
25989 where can i find mortgage rates
26070 precision engineering instruments
26075 where can i find html editors?
26161 where can i find an online translator?
26360 where can i download winnie the pooh and tigger
26465 how do i find someone's phone number
26487 how can i send flowers?
```

Figure 2: A sample of the judged queries used in the Large Web task.

Table 6: All official runs submitted in the Large Web task, presented by group. Groups are ranked in order of decreasing average P@10 of their best run. The number of groups was 6 and the number of runs "submitted" was 15. Unofficial runs inserted by the organisers (see text) are marked with an asterisk.

Group	Run tag	P@1	P@5	P@10
Microsoft*	ok8v1	0.5000	0.5214	0.5083
AT&T	att00100gb	0.5357	0.5190	0.4964
AT&T	att00100glf	0.5476	0.5048	0.4738
ACSys*	acsys9lnkA	0.4524	0.4643	0.4583
ACSys*	acsys9lnkE	0.5119	0.4929	0.4905
ACSys*	acsys9pr	0.3452	0.3167	0.2762
U Waterloo	uwmt9w100g0	0.4881	0.4500	0.4536
U Waterloo	uwmt9w100g1	0.4167	0.4262	0.4274
U Waterloo	uwmt9w100g2	0.4643	0.4548	0.4274
Fujitsu Labs	Flab9bN	0.4405	0.4619	0.4452
Fujitsu Labs	Flab9bsN	0.4524	0.4595	0.4381
Fujitsu Labs	Flab9rN	0.4405	0.4548	0.4440
Hummingbird	hum9w1	0.3095	0.3262	0.3250
Hummingbird	hum9w2	0.3095	0.3143	0.3167
Hummingbird	hum9w3	0.2857	0.3024	0.2940

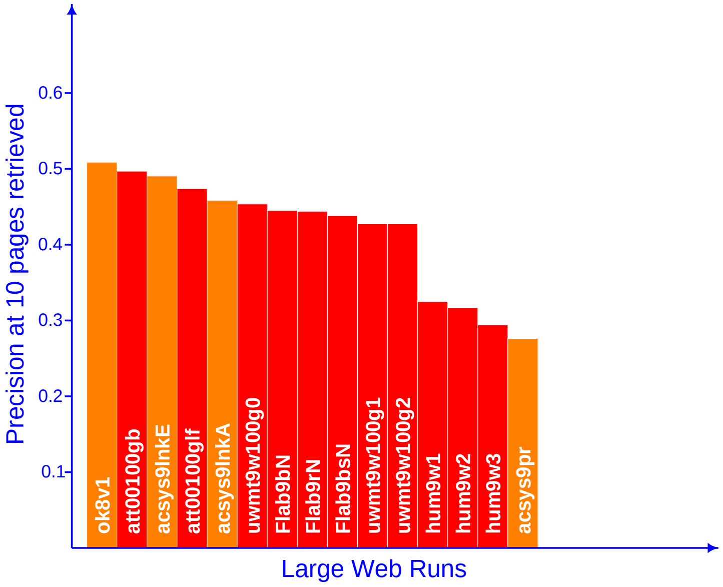

Figure 3: P@10 results for Large Task runs. Runs corresponding to the lighter coloured (orange) bars were submitted by the organisers and should not be compared with the other runs.

based on multiple levels of relevance, including identification of best pages. This finding is specific to a topic relevance task and is considered unlikely to apply to other forms of search task.

In the online service location task using the 100 gigabyte VLC2 collection, use of anchor text enabled AT&T to retrieve one more good document at rank one but otherwise no benefit was demonstrated on this task either. However, further work on this task is needed as the number of runs was small and there was little opportunity for tuning.

Useful information regarding the differences between Web and other TREC data has been accumulated through this year's Web track. Hopefully this will lead to increased participation in TREC-2001 and the use of better tested code.

Acknowledgements

With assistance from her colleagues at NIST, Ellen Voorhees played a major role in organising the Main Web part of the track, through topic formulation, assessment, evaluation and analysis. Much of the Main Web data and many of the analyses reported here are the result of her work.

The pivotal contributions of Peter Bailey and Nick Craswell in engineering the WT10g and preparing connectivity and other data are gratefully acknowledged.

We are very much indebted to Brewster Kahle of the Internet Archive for making available the spidered data from which the VLC2 and WT10g collections are derived and to Alta Vista (Monika Henzinger and Michael Moricz), eXcite (Jack Xu) 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 the NIST assessors for the Main Task assessments and to Sonya Welykyj, Penny Craswell, Julie Lemon, and Andrew Duncan for their work in assessing Large Task submissions.

Bibliography

Bailey, P., Craswell, N., and Hawking, D. 2001. Engineering a multi-purpose test collection for

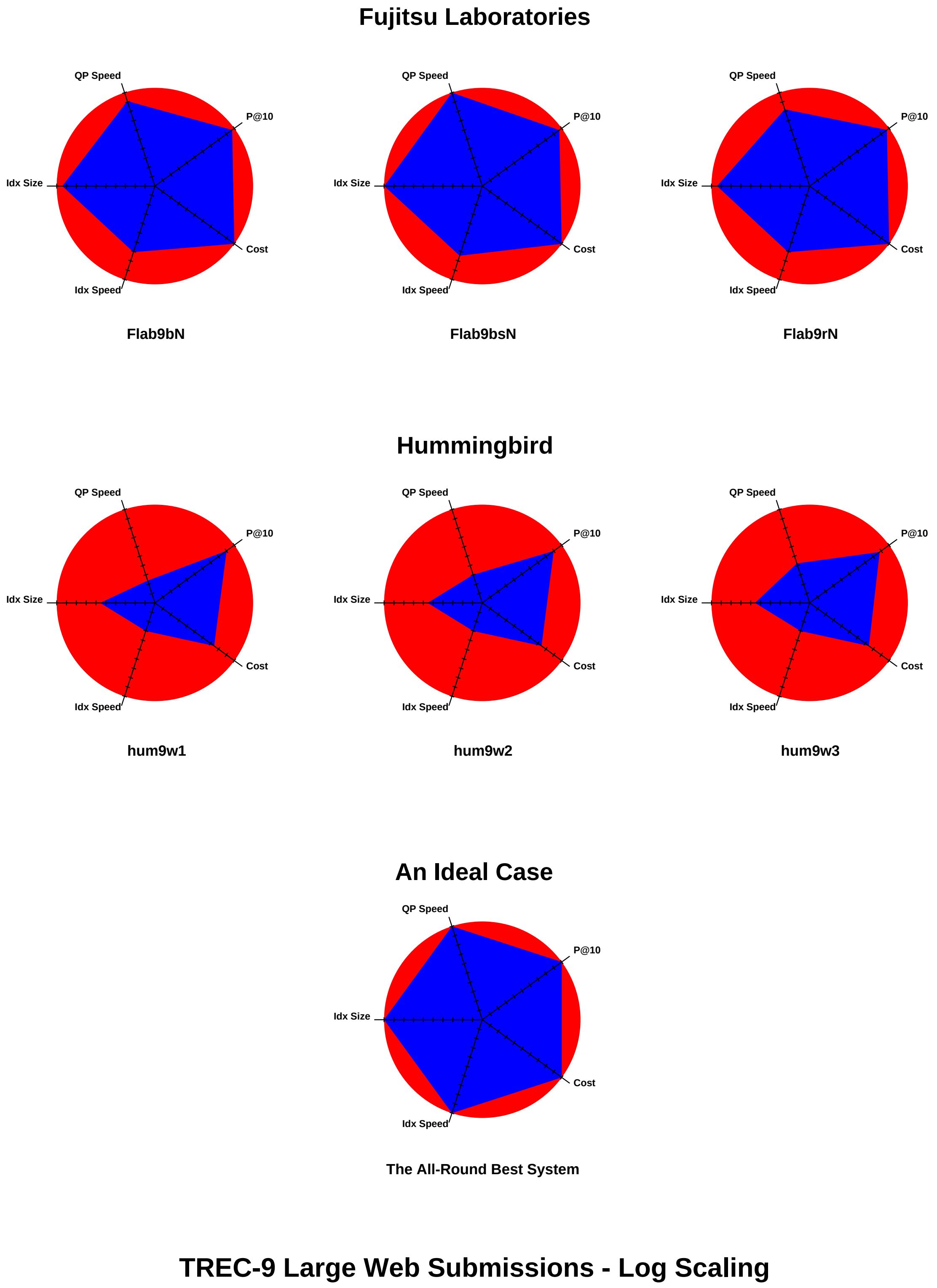

Figure 4: The trade-off between cost, speed, space and effectiveness for the runs for which questionnaire responses were received. Index size, indexing time and query processing times are scaled relative to the best values known to have been achieved in either TREC-8 or TREC-9.

web retrieval experiments. *Information Processing and Management*. In press. www.ted.cmis.csiro.au/~dave/cwc.ps.gz.

Craswell, N., Hawking, D., and Griffiths, K. 2001. Which search engine is best at finding airline site home pages? Technical Report XXX, CSIRO Mathematical and Information Sciences. www.ted.cmis.csiro.au/~nickc/pubs/airlines.pdf.

Hawking, D., Craswell, N., Bailey, P., and Griffiths, K. 2001. Measuring search engine quality. *Information Retrieval 4*, 1, 33–59.

Hawking, D., Craswell, N., and Griffiths, K. 2001. Which search engine is best at finding online services? In *WWW10 Poster Proceedings* (May 2001). www.ted.cmis.csiro.au/~dave/www10poster.pdf.

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), pp. 91–104. http://trec.nist.gov/pubs/trec7/t7_proceedings.html.

Järvelin, K. and Kekäläinen, J. 2000. Ir methods for retrieving highly relevant documents. In *Proceedings of SIGIR'00* (Athens, Greece, 2000), pp. 41–48.

Lynx. Lynx browser home page. http://lynx.browser.org.

Travis, B. and Broder, A. 2001. Web search quality vs. informational relevance. In *Proceedings of the 2001 Infonortics Search Engines Meeting* (Boston, 2001). www.infonortics.com/searchengines/sh01/slides-01/travis.html.

Voorhees, E. 2001. Evaluation by highly relevant documents. In *Proceedings of SIGIR'01* (New Orleans, LA, 2001). To Appear.

Zobel, J. 1998. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of ACM SIGIR'98* (Melbourne, Australia, August 1998).