Overview of TREC-7 Very Large Collection Track*

David Hawking
CSIRO Mathematics and Information Sciences,
Canberra, Australia
Nick Craswell and Paul Thistlewaite
Department of Computer Science, ANU
Canberra, Australia
David.Hawking@cmis.csiro.au, {pbt,nick}@cs.anu.edu.au

January 22, 1999

Abstract

In line with the wishes of last year's participants, this year's VLC track was essentially a re-run of last year's with a five-fold increase in data size. The data used was a completely new 100-gigabyte collection of Web documents (the VLC2) whose characteristics are presented here. This time, two orders of magnitude scale-up was investigated using 1%, and 10% samples as well as the full collection. Six groups managed to complete the full VLC task, of which five completed last year's track. An overview is given of the track participants, the methods used and the results obtained. One group of participants, using hardware costing less than $US10,000, have shown that a hundred gigabyte collection can be indexed in less than ten hours and that quite good rankings (better than several well-known search engines) can be produced from queries processed in less than one second.

1 Background and Motivation

The arguments for test collection sizes representative of the data sizes encountered in practice have been made by Harman [1992], Hawking and Thistlewaite [1997] and Hawking, Thistlewaite, and Harman [1999]. Within the last eighteen months, Web search engines have crossed the one hundred gigabyte data size barrier and some now closely approach the terabyte level. [Digital Equipment Corporation 1998]

At the same time, many TREC-6 VLC track participants expressed confidence that their systems were capable of indexing and querying collections much larger than 20 gigabytes.

Accordingly, a new 100-gigabyte test collection (the VLC2) has been developed for this year's track. The data used is part of a "Web-crawl" carried out by the Internet Archive in 1997. [Internet Archive 1997]

Naturally, as stated in last year's track overview, it is not feasible to obtain complete relevance judgments for collections of this size. Because of this, effectiveness measures are restricted to those which

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

can be derived from short rankings. (Fortunately, this year, it was feasible to judge a much larger number of runs than was the case last year.)

It was envisaged that TREC participants could examine in detail the effectiveness of their system on the main Ad Hoc task and then, if interested in larger collections, check speed and scalability in the VLC track. The VLC early precision measure exists mainly to ensure that speed is not achieved at the expense of effectiveness.

2 Organisers and Participants

As in the past, the VLC track was organised by the Advanced Computational Systems Cooperative Research Centre (ACSys), whose core participants are the Australian National University, the Commonwealth Scientific and Industrial Research Organisation, Fujitsu, Sun, DEC, StorageTek and Silicon Graphics. Support for the VLC track is a natural extension of ACSys research interests in "managing the information explosion".

ACSys obtained the tapes from the Internet Archive and supplied the human and machine resources to format and distribute the data. It also recruited and employed the VLC assessors. This year, a number of participating groups made financial contributions to the non-trivial cost of tape media and distribution.

Eleven groups received VLC2 data tapes. In the end, seven groups submitted runs: ACSys; City; UMass; UWaterloo; AT&T; the Okapi Group and FS Consulting. Three may be considered commercial organisations, three are universities and one is a government sponsored collaborative research centre.

3 The Data

Additional information on both editions of the Very Large Collection (VLC and VLC2) is available on the VLC web page. [Hawking et al. 1997]

A subset of the data tapes supplied by the Internet Archive was selected and formatted as the 100.426 gigabyte VLC2. From it, uniform 1% (BASE1) and 10% (BASE10) samples were defined as baselines.

The data was distributed on tape using `gzip` compression. The additional compression achieved by `gzip` (compared to standard Unix compress) saved considerably on tape cost and tape writing times. Even more effective compression systems are available but were not used, either because tests showed they would be too slow (up to eight days to compress the data and up to five days to decompress it on a Sun Ultra!) or because of the risk that some participating groups might experience difficulties in obtaining the necessary decompression code.

Complete sets of tapes were shipped to registered participants starting on June 15, 1998, allowing roughly eleven weeks to work on the task up to the submission deadline of September 8.

3.1 Access to the VLC Data

Access to the data is subject to the terms and conditions of the data permission forms available via the VLC Web page. [Hawking et al. 1997] These agreements prevent further redistribution, restrict use of the data to the usual TREC purposes and require recipients to delete documents if requested to do so by copyright holders, ACSys or the Internet Archive.

As previously mentioned, many of the groups participating in the track contributed to the cost of tape media and distribution, but financial contribution was not compulsory.

3.2 Overview of Data

The average document length is 5.67 kB compared to 3.2 for CDs 1-5 and 2.8 for the first edition VLC. The longest VLC2 document occupies about 4 MB. By comparison, the longest document (in the FR94 collection on CD4) in previously distributed TREC data was 6.2 MB.

The 1% and 10% baseline samples were created by selecting every 100th and every 10th compressed file respectively. BASE1 was thus a uniform sample of BASE10. Average document lengths in the samples are within a few percent of that for VLC2.

3.3 Formatting

The software used by the Internet Archive for "spidering" (collecting the pages from the Web) is an in-house system whose details are unknown to the VLC organisers. ACSys used `perl` scripts to convert the supplied data into VLC2 format. These scripts did not remove or convert any page content, merely inserting `<DOC>`, `<DOCNO>`, `</DOC>`, and `</DOCNO>` tags and a unique TREC-style document identifier. In addition, the HTTP header information (an average of 277 bytes) returned by the `httpd` daemon supplying the page, was surrounded by a `<DOCHDR>` `</DOCHDR>` pair. All pages with MIME type "text/html" were included, except a few longer than 2 MB.

It should be noted that this means that the collection includes some pages which are not in the English language, pages which are not in a Roman character set and pages which are in fact not text at all. (Some binary files (GIF files and compressed `tar` files) were erroneously typed by the daemons which served them). It also contains large numbers of duplicate (or near-duplicate) pages and pages which contain no text content (or very little).

Consequently, the VLC2 data is representative of the raw results likely to be obtained by Web spidering and thus representative of the pages likely to be indexed by Web search engines.

The 100.426 gB VLC2 is divided up into 97 collections, each in its own directory. Each collection is written as a separate tar file on tape. Collections each contain about 11 subdirectories, each containing around 48 bundles of documents (gzipped files), each containing on average 370 documents.

It has been distributed in three different tape formats: DLT-4000 (2 tapes, second includes baselines), DDS-3 (three tapes for VLC2, fourth for baselines) and DDS-2 (eight tapes for VLC2, ninth for baselines).

VLC2 document identifiers are structured to allow unambiguous identification of collection, subdirectory and filename. Every document contained the essential "SGML" markers delimiting documents and document identifiers. A program `coll_check` was used to check that each document conformed to this elementary structure and that document identifiers were unique. The only problem detected (initially) by this program was caused by a fragment of a TREC Federal Register document appearing on someone's Web site!

4 The Task

Full guidelines for the VLC track are available on the VLC web page [Hawking et al. 1997]. In essence, participants were required to process queries generated from the TREC-7 Ad Hoc topics (351-400) over both the baselines and the VLC2 datasets and to return for assessment only the first 20 documents retrieved in each case. Elapsed times (as would have been observed by a human with a stopwatch) for indexing the datasets and processing queries were recorded and system details and costs as well as disk space requirements were reported via a questionnaire. The focus was on the ratios of the various measures (see below) across the three data sizes

All retrieved documents were judged.

Participants were encouraged to submit at least one set of results using queries derived automatically from the Title and Description fields of the topic statements.

5 The Measures

M1. Completion. (Can the system process data of this size at all?)

M2. Precision@20. (P@20)

M3. Query response time. (Elapsed time as seen by the user.)

M4. Data Structure Building time. (Elapsed time as seen by the user.)

M5. Gigabyte-queries/hour/kilodollar. (Bang per buck.)

M6. Modified average precision. This is a new measure introduced to take account of the fact that, for some topics, the number of relevant documents in a collection (BASE1, BASE2 or VLC2) may be so small as to artificially limit P@20. M6 is calculated by summing the precision at each point in a ranking where a relevant document occurs and dividing the sum by the lesser of 20 and the total number of relevant documents for the topic in that collection.

M4 represented the minimum possible elapsed time from receiving the data until the data structures necessary to process the queries used in M3 were built, using the chosen hardware and indexing software. Time to actually read the tapes was excluded. The starting point was the compressed data files on disk after unpacking the tarfiles. M4 included the time to build all structures (such as inverted files) which are necessary to process the final query.

6 The Assessments

Four judges were employed to assess the VLC2 document pool. One was a research assistant in Sociology, another a final year Philosophy/Art Curatorship student with employment experience in summarisation of technical articles, another a Science graduate and the fourth a graduate in both Arts/Asian Studies and Science. The first judge was also employed in the TREC-6 track.

Topics were assigned to judges on an arbitrary basis. All judgments for a particular topic were made by the same judge.

Groups were permitted to submit multiple sets of runs but were asked to indicate a priority order for assessment. As it turned out, ALL runs submitted prior to the deadline were assessed and the resulting `qrels` and evaluations were distributed on October 14th.

When it became clear that good progress was being made on judging, groups were offered the chance to submit additional runs (or deeper rankings for previously submitted runs.) These after-deadline runs were all completely judged. Unfortunately, due to restrictions on the availability of judges, it was necessary to transfer responsibility for some topics from one judge to another. In these cases, the new judge re-judged all the before-deadline documents on those topics and the old judgments were discarded. The result is that there are two different sets of qrels. In both sets, only one judge was used per topic.

Table 1: Groups completing the VLC task. Six groups attempted the full 100 gigabyte task and one additional run was submitted using just the BASE1 collection. The hardware configuration shown is the full configuration available. In some cases, groups used only part of the available configuration, even on the 100 gigabyte task and in some cases, groups used less hardware on the baselines. The pair of figures in the I/O column indicate the number of channels and the number of disks used (per CPU, unless otherwise noted). Cost is an estimate of the U.S. list price of a system comparable to the one used.

Group	Software	CPUs	MHz	Total RAM	I/O	Cost
ACSys	PADRE98	8 x DEC Alpha	266	1152MB (dist.)	1,2	$24k
ATT	Smart	20 x SGI R10000	195	8192MB (sh.)	1,1	$115k
City	PLIERS	8 x DEC Alpha	266	1152MB (dist.)	1,2	$24k
FSC	MPS	1 x Sun Ultra	200	256MB	3,19	$15k
Okapi	Okapi	2 x Intel P2 (Solaris)	400	512 MB (sh.)	?,15(tot.)	$37k
UMass	Inquery	4 x Sun Ultra	167	1024 MB (sh.)	?,?	$130k
Waterloo	Multitext	4 x Intel P2	300	512 MB(dist.)	2,4	$8k

Table 2: M2: Precision at 20 documents retrieved. Numbers in parentheses represent ratios to the appropriate baseline measures. The last column characterises the method used to generate this query set. T, D, and N refer to the Title, Description and Narrative fields of the topic. RF refers to automatic relevance feedback, and *Cov. dens.* refers to cover density ranking. *Req. wds.* indicates that automatically generated required words were added to the query.

Group	BASE1	BASE10	VLC2	Q gen.
UMass(1)	.202	.429(2.12)	.625(3.09,1.46)	T+D+N RF
UMass(2)	.204	.441(2.16)	.624(3.06,1.42)	(1) + Req. wds.
UMass(3)	.208	.419(2.01)	.598(2.88,1.43)	T+D RF
Okapi(1pr)	.180	.376(2.09)	.541(3.01,1.44)	T+D
Okapi(1pr.tnd)	-	-	.598(3.32,1.59)	T+D+N
Okapi(3)	-	-	.509(2.83,1.35)	T+D RF
Okapi(3.tnd)	-	-	.545(3.03,1.45)	T+D+N RF
Waterloo(0)	.190	.369(1.94)	.442(2.33,1.20)	T Cov. dens.
Waterloo(1)	.235	.474(2.02)	.598(2.55,1.26)	Manual (6.4 term)
Waterloo(2)	.223	.411(1.84)	.574(2.57,1.40)	Manual (1.9 term)
Waterloo(3)	.110	.288(2.62)	.397(3.61,1.38)	Variant of (0)
ATT (vf)	.188	.384(2.04)	.503(2.68,1.31)	T+D
ATT (vfe)	-	-	.587(3.12,1.53)	T+D RF
ATT (vi)	-	-	.357(1.90,.930)	T+D
ATT (vie)	-	-	.375(1.99,.977)	T+D RF
ACSys (5)	.139	.321(2.31)	.442(3.18,1.38)	T+D (5 term)
ACSys (2)	-	-	.298(2.14,.930)	T+D (2 term)
FSC	.128	.268(2.09)	.345(2.70,1.29)	T
City (1)	.080	-	-	T+D
City (2)	.056	-	-	T+D

Table 3: M3: Average Query Processing Time (Elapsed seconds per query). Numbers in parentheses represent ratios to the appropriate baseline measures.

Group	BASE1	BASE10	VLC2
ACSys (5)	0.061	0.168(2.75)	1.47(24.1,8.74)
ACSys (2)	-	-	0.887(14.5,5.28)
ATT (vf)	1.44	6.41(4.45)	5.80(4.03,0.906)
ATT (vfe)	-	-	12.0(8.33,1.87)
ATT (vi)	-	-	2.18(1.52,0.341)
ATT (vie)	-	-	8.00(5.58,1.25)
City (1)	0.593	-	-
City (2)	1.74	-	-
FSC	0.10	0.46(4.6)	51.8(518,113)
Okapi (1pr)	0.96	3.74(3.88)	25.9(26.9,6.92)
Okapi (1pr.tnd)	-	-	81.5(84.5,21.8)
Okapi (3)	-	-	68.0(70.6,18.2)
Okapi (3.tnd)	-	-	105(109,28.1)
UMass (1)	8.4	85.2(10.2)	712(84.7,8.35)
UMass (2)	9.6	88.8(9.25)	718(74.7,8.07)
UMass (3)	7.2	54(7.5)	526(73,9.73)
Waterloo (0)	0.306	0.294(0.960)	0.708(2.31,2.41)
Waterloo (1)	0.216	0.377(1.75)	1.51(6.99,4.00)
Waterloo (2)	0.251	0.299(1.19)	0.882(3.51,2.95)
Waterloo (3)	0.148	0.212(1.43)	0.619(4.18,2.92)

Table 4: M4: Data Structure Building Time (Elapsed Hours). Numbers in parentheses represent ratios to the appropriate baseline measures. UMass indicated that the starred times are likely to be significant over-estimates of the true values.

Group	BASE1	BASE10	VLC2
ACSys	0.0434	1.71(39.4)	7.73(178.1,4.52)
ATT	0.43	5.14(12.0)	6.55(15.2,1.27)
City(1)	0.0794	-	-
FSC	1	10(10)	100(100,10)
Okapi	0.42	3.85(9.167)	36.1(86.0,9.38)
UMass	2.67*	26.15(9.79)*	35.45(13.28,1.36)
Waterloo	0.0519	0.504(9.71)	5.33(102.7,10.6)

Table 5: MS: Data Structure Sizes (gigabytes). Numbers in parentheses represent ratios to the appropriate baseline measures. In the case of Okapi, the size of the raw text (compressed) must be added for queries which use relevance feedback. The UMass system also expects the raw text to be available.

Group	BASE1	BASE10	VLC2
ACSys	0.255	0.902(3.53)	7.70(30.2,8.54)
ATT	0.329	2.09(6.35)	20.8(63.2,9.95)
City(1)	0.124	-	-
FSC	0.27	2.5(9.26)	24(88.9,9.6)
Okapi	1.58	11.5(7.28)	109.7(69.4,9.54)
UMass	0.68	6(8.82)	53(77.9,8.83)
Waterloo	0.789	6.25(7.92)	44.6(56.5,7.14)

Table 6: M5: Gigabyte-queries per hour per kilodollar for 100 gigabyte runs. For each group the fastest run of queries derived automatically from T+D fields is presented.

Group	Runid	Queries/Hr	kilo$	gB-Q/Hr/kilo$	Last Year
Waterloo	uwmt7v3	5816	8.5	6.84×10^4	7.20×10^3
UMass	inq5vlc3	6.84	130	5.26×10^0	3.8×10^0
ATT	att98vi	1651	115	1.44×10^3	7.89×10^1
FSC	fsclt7a-v100	69.5	15	4.63×10^2	NA
Okapi	ok7vf1pr	139	37	3.76×10^2	6.88×10^1
ACSys	acsys7_100_2	4059	24	1.69×10^4	1.50×10^1

Table 7: M6: Modified average precision. Numbers in parentheses represent ratios to the appropriate baseline measures.

Group	BASE1	BASE10	VLC2
ACSys (5)	.1535	.2326(1.52)	.3108(2.03,1.34)
ATT (vf)	.2098	.3035(1.45)	.3810(1.82,1.26)
FSC	.1490	.1975(1.33)	.2407(1.62,1.22)
Okapi (1pr)	.2155	.2812(1.31)	.3957(1.84,1.41)
UMass (1)	.2463	.3407(1.38)	.5201(2.11,1.53)
UMass (2)	.2677	.3514(1.31)	.5143(1.92,1.46)
UMass (3)	.2485	.3316(1.33)	.4832(1.94,1.46)
Waterloo (0)	.2472	.2897(1.17)	.3380(1.36,1.17)
Waterloo (1)	.3099	.3902(1.26)	.5005(1.62,1.28)
Waterloo (2)	.2728	.3360(1.23)	.4659(1.71,1.39)
Waterloo (3)	.1428	.2179(1.53)	.2921(2.05,1.34)

References in the present paper to pool sizes and relevant sets relate to the before-deadline submissions and judgments. Groups wishing to compare new runs with official runs reported here, must re-score the official runs using the new qrels in `http://pastime.anu.edu.au/TAR/Qrels/`.

The document pool (derived from baseline *and* VLC submissions) contained 16,292 document/topic pairs of which 4,440 were judged relevant. By contrast, the corresponding figures for the TREC-7 Ad Hoc task (using the same topics but a disjoint set of documents) were 80,345 and 4,674.

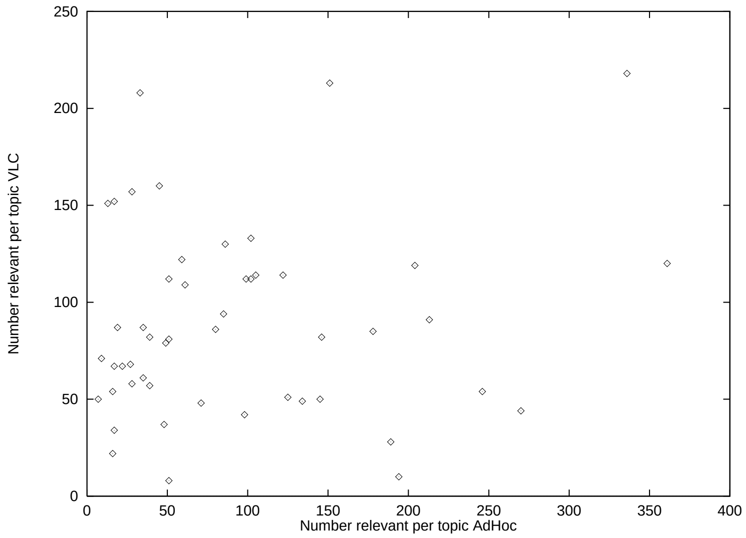

Figure 1: The relationship between numbers of known relevant documents found for the same 50 topics in two disjoint corpora: the VLC2 and the TREC-7 Ad Hoc corpus. VLC2 judgments used were the "before-deadline" set. Pearson $r = 0.13$.

The range of numbers of relevant documents found in the VLC2 (including the baselines) per topic was 8 - 218 (compared to 7 - 361 for the Ad Hoc task.) The figure of 218 is almost certainly an underestimate due to the small number of runs and the shallow judging.

Each point in the scatter plot in Figure 1 plots the number of relevant documents for the VLC2 collection against the corresponding number for the Ad Hoc collection for a particular topic. There is no significant correlation between the number of relevant documents in the two collections (Pearson $r = 0.13$, $p > 0.05$).

In Figure 2, the topics have been ordered by increasing number of relevant documents: a) for the Ad Hoc task; and b) for the VLC (using both before and after-deadline judgments. The number of relevant documents has been plotted against topic rank for each ordering. Considering the before-deadline judgments, the number of relevant documents per topic is generally greater for the VLC2 than for the Ad Hoc collection up to about 85 documents found relevant, after which the contrary is the case. It is possible that this may be the incompleteness of the VLC2 judgment pool starting to manifest itself. Indeed, when using the more complete after-deadline VLC judgments, the VLC2 line remains above the Ad Hoc line up to about 180 documents found relevant.

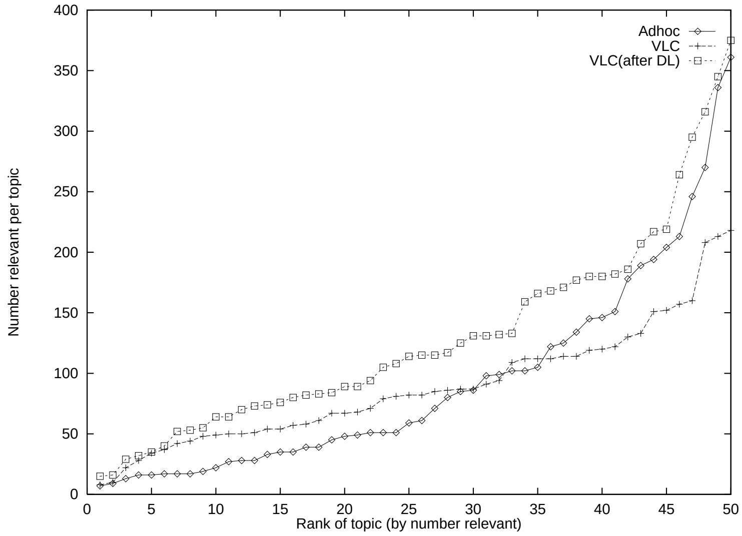

Figure 2: Ordered rankings of number of relevant documents per topic for the same 50 topics in two disjoint corpora: the VLC2 and the TREC-7 Ad Hoc corpus. Two traces are shown for the VLC2 corpus, one using "before-deadline" judgments and the other the revised set incorporating judgments of runs submitted after the deadline.

6.1 Were the Baseline Collections Unbiased Samples?

This is an important question, because it may determine the "scalability" of early precision and perhaps influence other measures. The process of selecting the baseline subsets (described above) is not inherently biased but the results may be biased with respect to a particular set of topics.

Of the 7445 documents retrieved in the runs over the full VLC2, 491 were also in BASE10 and 48 were also in BASE1. The proportion of documents in the VLC2, BASE10 and BASE1 which were retrieved by VLC (not baseline) runs were 4.01×10^{-4}, 2.64×10^{-4} and 2.56×10^{-4}, respectively. Unfortunately, tests of one-sample proportion (with finite sample correction) show that, for both baselines, the sample proportions lie outside the 95% confidence interval. Hence, it appears that samples are biased with respect to proportion of retrieved documents. Consequently, caution must be exercised when assessing increases in precision from baselines to the full collection.

By contrast, the same test applied to the BASE10 pool (4,500 documents) gives no reason to suggest that BASE1 is a biased sample of the BASE10 collection. (The proportions were 2.42×10^{-3} v. 2.27×10^{-3}).

6.2 Baseline Relevant Sets

There were 497 documents judged relevant in the BASE1 collection. There were three topics for which no relevant documents were found in BASE1.

There were 1673 documents judged relevant in the BASE10 collection. There was only one topic for which no relevant documents were found in BASE10.

7 Characteristics of Submitted Runs

The seven groups which submitted runs are listed in Table 1.

7.1 Hardware Used

Three shared-memory systems were used: ATT's 20-processor Silicon Graphics system, UMass's 4-processor Sun SPARCserver, and Okapi's 2-processor Dell. No group completed the 100 gigabyte task using a single procesor, however Okapi indicated that the second CPU made very little difference to elapsed times in their case.

Note that:

1. A majority of groups used hardware they had access to rather than explicitly choosing it for the task. Their systems may have run just as fast on much cheaper hardware.

2. Few groups were able to run their system in dedicated mode. It is difficult to control for the effect of other users.

3. It is difficult to derive a comparable dollar value for a fraction of a very expensive system or for obsolete systems.

8 Questions Addressed

The main questions addressed by participants basically related to demonstrating the capabilities of their system on the large scale task, measuring speed and effectiveness and also observing the scalability of their systems.

Other questions will presumably be covered by each group in their own paper.

9 The Results

The results are presented in Tables 2-6

1. Table 2 shows that there is a significant rise in P@20 for all systems moving from the sample collections to the full VLC2. This is consistent with observations in last year's VLC track ([Hawking and Thistlewaite 1997; Hawking et al. 1999]) but is confounded by the evidence that the samples appear to have been biased with respect to the particular set of topics. It is almost certainly the case that a major part of the explanation of the precision increase is that precision in the small samples is severely limited by the number of relevant documents in the sample for many topics. Even a perfect retrieval system cannot achieve non-zero P@20 if there are no relevant documents.

2. Table 7 shows that modified average precision decreases less dramatically than P@20 but still increases. Note that nearly all runs included topics which failed to find any relevant documents and consequently scored zero on modified average precision. Modified average precision was zero for

between 6 and 25 topics (median 13) on the BASE1 runs and for between 1 and 7 topics (median 4) on the BASE10 runs. Even on the 100 gigabyte task, there was only one run `uwmt7vi` which found relevant documents on every topic. The number of topics scoring zero on modified average precision ranged from 0 to 8 (median 2) for the VLC runs.

3. Table 3 shows that two groups (ACSys and Waterloo) achieved sub-second query processing time over the full collection. In the UMass case, query processing time was approximately proportional to the corpus size, while for FSC, query processing time grew non-linearly, presumably due to virtual memory effects. ACsys, ATT, Okapi and Waterloo were able to control dilation of query processing time to a factor less than the collection scale-up. In general, this is because when the data-dependent parts of query processing are made more efficient, fixed costs (such as accessing 20 document identifiers and looking up terms in a term dictionary) become significant for small collections. In the ATT case this was also partly due to the use of more hardware for the larger collections.

4. Table 4 shows that, for FSC, Okapi and Waterloo, data structure building time is an essentially linear function of data size. This may have been the case for UMass too, but they have indicated that their baseline timings are not reliable. ATT controlled the increase in running time by deploying more hardware for the larger collections. ACSys used the same hardware for each collection but the running time of its algorithm is heavily dependent both upon the balance between the size of the chunk of text being indexed and the available RAM and upon load balance between workstations. With more RAM, and more uniform load balance the relationship may have been more linear.

5. Table 5 shows considerable variation in index size across the groups, ranging from 8% (ACsys) to 110% (Okapi) of the raw data size. All groups observed a scale-up in index size of between 7 and 10 for the transition from BASE10 to VLC2. However, ACSys observed a much lower scale-up for the transition from BASE1 to BASE10. This is believed to be because the uncompressed term dictionaries are larger relative to the raw data size in the BASE1 case (with many entries replicated over 8 workstations).

6. Table 6 shows that “bang-per-buck” measures have improved considerably since last year.

10 How Would Commercial Web Search Engines Perform?

Table 8: M2/M6: P@20 and modified average precision performance for Web Search Engines, using title-only queries and the real Web. The range of official VLC results (re-evaluated using after-deadline judgments to ensure comparability) is also shown.

Engine	1	2	3	4	VLC range	VLC median
P@20	.306	.288	.231	.377	.283 - .617	.490
Mod. Ave prec	.2228	.2000	.1262	.2693	.1535 - .5154	.3736

Out of interest, TREC-7 title-only queries were fed to four well-known Web search engines. The engines were searching the current Web rather than the VLC2 frozen snapshot. Top 20 results for each of the topics over the real Web were then presented to the same judge who had judged the documents from VLC2 for that topic, using the same assessment interface and the same concepts and evidence that they had built up.

Results for these search engines are presented in Table 8. As may be seen, all four engines perform well below the median for VLC submissions on both P@20 and modified average precision.

11 Discussion and Conclusions

1. Considerable speed improvement has been achieved by most of the five groups who participated last year. Apart from additional disk storage, the gain was achieved without massive hardware upgrades. In many cases, the appropriate cost to assign to systems used actually declined, either because expensive hardware was not used or because old machines dropped off vendor price-lists and comparable performance is now available much more cheaply.

2. It should be no surprise, given popular experience with large Web search engines, that sub-second query processing is possible over collections the size of the VLC2. What is perhaps surprising is that such performance is possible from relatively small scale hardware.

3. In the absence of huge amounts of RAM, query processing using uncompressed indexes comes to be dominated by disk latencies. For example, if the list of document names (of the order of 300 MB uncompressed) is not memory-resident, then 20 disk accesses at 10-15 msec. each (0.2 - 0.3 sec.) are likely to be needed to produce a top 20 ranking. Accepting that there must be at least one I/O request per query term to retrieve the posting list, it is important to minimize the number of I/O requests required to *locate* the posting list and also important to avoid dividing the collection into multiple sub-collections which must be separately searched.

4. It would be possible to speed up query processing dramatically if huge amounts of RAM were available. For example, in the ACSys case, a total of 8 gigabytes of RAM (as used in each of the current Alta Vista query processing engines) would be sufficient to load ALL data structures for the 100 gB collection into memory, totally obviating the need for disk I/O. Even half of this would suffice for most queries and compression techniques would reduce RAM requirements still further.

5. The effectiveness scores of the public Web search engines are all considerably below the median for the VLC participants, despite the fact that some of them have access to many more documents. Some VLC participants used very much longer queries (and two runs were manually generated), but P@20 results for the two Title-only runs submitted by VLC participants were still better than the best Web engine tested.

12 The ACSys VLC Medal

ACSys offered a medal to any group submitting a 100-gigabyte run which achieved:

1. Average query processing speed of 2 seconds or less;

2. Indexing time under 10 hours.

3. Median P@20 or better.

Three groups (ATT, Waterloo and ACSys (ineligible for the medal)) achieved the first two criteria. ATT also met the third criterion with its vfe run but unfortunately this run did not satisfy the first criterion.

Waterloo met all three criteria with two separate runs, one of which processed queries in an average of less than 0.9 seconds. Both of these query sets were classified as Manual runs but no query generation method restrictions were stated in the medal conditions.

Considering only 100 gB runs which were automatically generated from no more than the T+D fields, six runs were excluded and the median P@20 dropped from 0.525 to 0.442. This left Waterloo run 0 and the ACSys 5-term run exactly on the median.

Accordingly, Waterloo was presented with an ACSys VLC medal during the VLC plenary session at TREC-7.

Acknowledgements

We are very much indebted to Brewster Kahle of the Internet Archive for making available the spidered data (and trusting us with a difficult-to-replace set of tapes) and to Edward King of the Earth Observation Centre for sparing us a considerable amount of his time and expertise in converting tape formats. Thanks also to Donna Harman and Ellen Voorhees of NIST, to John O'Callaghan (CEO of ACSys) and to the TREC Program Committee for supporting the track.

Finally, thanks are due to Sonya Welykyj, Penny Craswell, Nick Clarke and Angela Newey for good work in assessing submissions.

Bibliography

Digital Equipment Corporation. 1998. Digital's Alta Vista search index grows to record heights. `http://www.altavista.digital.com/av/content/pr052798.htm`. Press release.

Harman, D. K. Ed. 1992. *Proceedings of the First Text Retrieval Conference (TREC-1)* (Gaithersburg MD, November 1992). U.S. National Institute of Standards and Technology. NIST special publication 500-207.

Hawking, D. and Thistlewaite, P. 1997. Overview of TREC-6 Very Large Collection Track. In E. M. Voorhees and D. K. Harman Eds., *Proceedings of the Sixth Text Retrieval Conference (TREC-6)* (Gaithersburg MD, November 1997), pp. 93–106. U.S. National Institute of Standards and Technology. NIST special publication 500-240.

Hawking, D., Thistlewaite, P., and Craswell, N. 1997. *TREC Very Large Collection (VLC) web page*. ACSys Cooperative Research Centre, The Australian National University, Canberra. `http://pastime.anu.edu.au/TAR/vlc.html/`.

Hawking, D., Thistlewaite, P., and Harman, D. 1999. Scaling up the TREC Collection. Accepted by *Information Retrieval* September 1998.

Internet Archive. 1997. Building a digital library for the future. `http://www.archive.org/`.