Overview of TREC-6 Very Large Collection Track

David Hawking and Paul Thistlewaite
Co-operative Research Centre For Advanced Computational Systems
Department Of Computer Science
Australian National University
{dave,pbt}@cs.anu.edu.au *

January 10, 1998

Abstract

The emergence of real world applications for text collections orders of magnitude larger than the TREC collection has motivated the introduction of a Very Large Collection track within the TREC framework. The 20 gigabyte data set developed for the track is characterised, track objectives and guidelines are summarised and the measures employed are described. The contribution of the organizations which made data available is gratefully acknowledged and an overview is given of the track participants, the methods used and the results obtained. Alternative options for the future of the track are discussed.

1 Background and Motivation

In the overview of the proceedings of TREC-1, Harman [1992] referred to small early test collections such as Cranfield, CACM and NPL and argued the need for a realistically-sized test collection to facilitate the transfer of laboratory-developed retrieval systems into the field. The 2-gigabyte collection used in TREC-1 was two orders of magnitude larger than previous collections, and legitimately given the label of a *very large test collection*. Indeed, given the state of contemporary hardware and indexing software, it posed considerable challenges to participants.

Two gigabytes remains a realistically-sized test for text retrieval applications typical of universities, research organisations, newspapers, businesses and government departments. However, it is clear that some organisations such as patent offices and future digital libraries will demand retrieval services over collections at least two orders of magnitude larger, despite trends toward distributed information retrieval. There are already collections of the 100 gigabyte scale in the commercial world and Web search engines such as HotBot claim to index in excess of 50 gigabytes.

Accordingly, in line with the initial TREC charter of realism, a need was identified for a test collection significantly larger than that used in mainstream TREC. It was not intended to replace the main TREC collection but rather to be used in a special-interest Very Large Collection (VLC) track to allow interested developers of commercial and research retrieval systems to investigate the scalability of their methods. It would also help to verify that such systems did not suddenly cease to operate due to machine or operating system limits on virtual addressing, file system

*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.

size etc. and allow some effectiveness comparison of systems currently operating with very large collections. The proposed collection size was 20 gigabytes, a factor of ten larger than that used in the TREC mainstream task. This collection size seemed feasible, as hardware and software improvements since TREC-1 had dramatically reduced the difficulty of working with gigabyte-scale collections.

The value of a test collection lies not only in the data itself but in the availability of judgments of its documents as to relevance to a large set of research topics. Complete sets of judgments are available for some test collections but are not affordable for TREC-sized collections. (At an optimistic judging rate of 500 documents per judge per working day, complete judgment of a collection of one million documents requires about eight person-years *per topic*!) TREC approximates a complete set of judgments for its topics by manually judging only those documents in the pool retrieved by a [hopefully] diverse set of automatic retrieval systems and deeming that un-judged documents are irrelevant. This allows recall-oriented measures to be determined with a reasonable (but not perfect) degree of confidence.

Assessment resources available to the VLC track are not sufficient to support recall-oriented measures over a 20-gigabyte collection. Even if sufficient resources were available to support the TREC pooling method, that method is not likely to be effective in the VLC context. For any given topic there may be ten times as many relevant documents as in the standard TREC task yet the reduced number of participating systems is likely to mean that fewer are judged.

Accordingly, effectiveness measures in the VLC track were confined to the precision dimension. It was envisaged that TREC participants could demonstrate the speed and effectiveness merits of their system on the main AdHoc task and then, if interested in larger collections, demonstrate how speed was affected by a ten-fold increase in data size and (hopefully) confirm that speed results were not achieved at the expense of lost precision.

A trial run of the VLC track took place in TREC-5 (1996) using CDs 1-4 of the TREC set (a total of 4.28 gigabytes). Four groups submitted runs, judgments made by Canberra assessors were validated against those in Washington and various issues were clarified for the running of the track proper at TREC-6.

2 The Organisers

The VLC track (like the pre-track in TREC-5) has been 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 and DEC. Support for the VLC track is a natural extension of ACSys research interests in scalable computing and large datasets.

With full support from NIST and the TREC program committee, ACSys collected the additional data to make up the VLC and supplied the human, financial and machine resources to format and distribute the data. It also recruited and employed the VLC assessors.

3 The Participants

Fourteen groups, including 6 universities, received VLC data tapes. One registered very late and was unable to read the tapes. In the end, seven groups submitted runs, comprising four universities and three commercial groups: ANU/ACSys, City, UMass, UWaterloo, AT&T and IBM (two separate groups).

4 The Data

Additional information on the Very Large Collection is available on the VLC web page [Hawking et al. 1997].

A 20.14 gigabyte collection (including all five TREC CD-ROMs) was assembled with assistance from a large number of data holders. From it, a uniform 10% sample was defined for use as a baseline.

The additional (non-NIST) data was distributed on DAT (DDS-1) format tapes due to logistical and economic difficulties with using CD-ROMs. Participants reported some difficulties in reading these tapes but only in one case (a late starter) were these responsible for a non-submission. The final set of tapes was shipped to all registered participants (at the time) on June 20, 1997, allowing participants roughly nine weeks to work on the task up to the submission deadline of September 8.

4.1 Access to the VLC Data

Access to the data (except for USENET news data) is subject to the terms and conditions of the TREC data permission forms. Copyright owners only granted permission to distribute the data on this basis. These owners are listed in the Acknowlegements below. Permissions were obtained from controllers of all websites used as sources of documents.

4.2 Overview of Data

The VLC data is somewhat biased by the inclusion of roughly 8.7 gigabytes of USENET news postings to make up the target 20 gigabytes. This data has a significantly different character to the data on CDs 1-5. However, the remainder of the non-NIST data in the VLC adheres reasonably well to the earlier TREC pattern and represents a diversity of sources covering government agencies (eg. Australian Department of Industrial Relations), parliamentary proceedings (Canadian and Australian Hansards) and newspapers (eg. Glasgow Herald and Financial Times). For the first time, HTML documents downloaded from the Internet are included (eg. CSIRO and Australian university websites). Also for the first time, there is a large quantity of legal data including both laws and judgments, thanks to the Australian Attorney General's Department. The latter is mostly in HTML format.

Collections in the new VLC data are typically larger than those on CDs 1-5. However, addition of the new data has not altered the minimum or maximum document length figures. Average document length has declined slightly, from 3.2 kilobyte for CDs 1-5 to 2.8 kilobyte for the entire VLC.

The 10% baseline sample was created by selecting every 10th compressed file and then manually removing an arbitrary handful of files to bring the sample to a closer approximation of 10%. Average and minimum document lengths changed by negligible amounts but the longest document in the baseline dropped to 2.8 MB from 6.2 MB.

4.3 International Balance

The international balance of the data is significantly different to the combined NIST data, of which 90% is sourced in the U.S. Ignoring the NEWS and Project Gutenberg collections, whose origins are mixed but U.S.-dominated, the remaining 11.3 gigabytes is sourced roughly 41% from the U.S., 44% from Australia, 10% from England, 4% from Scotland and less than 1%

Table 1: Crude breakdown of VLC, VLC assessment pool and VLC relevant set by source.

Source	# Documents	# Documents judged	# Relevant documents
TREC6 docs.	556,077(7.4%)	1608(18.9%)	631(21.7%)
Other NIST docs.	1,078,166(14.4%)	3426(40.3%)	1202(41.3%)
All ACSys-collected docs.	5,857,805(78.2%)	3477(40.9%)	1076(37.0%)
- USENET news docs.	4,400,657(58.7%)	2001(23.5%)	552(19.0%)
- ACSys non-USENET docs.	1,457,148(19.4%)	1476(17.3%)	524(18.0%)

Table 2: Probability of retrieval and probability of relevance for documents from different sources. (Obtained by dividing the raw frequencies in table 1 by the number of documents from each source.) The last column gives the probability for each source that a document in the assessment pool is actually relevant.

Source	Pr(retrieved)	Pr(relevant)	Pr(relevant|retrieved)
TREC6 docs.	0.00289	0.00113	0.392
Other NIST docs.	0.00318	0.00111	0.351
All ACSys-collected docs.	0.000593	0.000184	0.309
- USENET news docs.	0.000455	0.000125	0.276
- ACSys non-USENET docs.	0.00101	0.000360	0.355

Table 3: Contributions of individual ACSys collections to the VLC pool and the VLC relevant set. The probability ratios are computed by calculating the probability that a document from this source will be part of the pool (or part of the relevant set) and dividing this by the corresponding probability for all NIST-collected documents.

	Collection		Pool			Relevant Set		
Source	MB	# docs	# docs	% of pool	Prob. Ratio	# docs	% of rel. set.	Prob. Ratio
AAG	1874.5	61,566	230	2.7%	0.133	59	2.0%	0.094
ADIR	775.0	42,841	9	0.1%	0.068	1	0.0%	0.021
APLT	1539.8	421,681	501	5.9%	0.386	185	6.4%	0.391
AUNI	724.8	81,334	134	1.5%	0.535	40	1.4%	0.438
FT	526.7	202,433	259	3.0%	0.415	100	3.4%	0.440
GH	393.6	135,477	251	2.9%	0.601	107	3.7%	0.704
NEWS01	954.5	446,106	180	2.1%	0.131	65	2.2%	0.130
NEWS02	943.1	450,027	221	2.6%	0.159	63	2.2%	0.125
NEWS03	936.6	482,395	228	2.7%	0.153	56	1.9%	0.104
NEWS04	966.0	83,145	233	2.7%	0.157	61	2.1%	0.113
NEWS05	1169.7	590,202	325	3.8%	0.179	91	3.1%	0.137
NEWS06	1120.6	571,891	260	3.1%	0.148	47	1.6%	0.073
NEWS07	1080.1	520,282	240	2.8%	0.150	60	2.1%	0.103
NEWS08	1727.9	856,609	314	3.7%	0.119	109	3.7%	0.113
PGUT	430.3	3,303	30	0.4%	2.949	5	0.2%	1.350
WEB01	141.9	8,513	62	0.7%	2.364	27	0.9%	2.828

from Canada. These proportions reflect the availability of data rather than any goal of the organisers. The proportion of non-English-language text in the VLC is negligible.

4.4 Formatting

A variety of `flex` programs and `perl` scripts were used to convert supplied data into VLC format. The `wget` program was used to download web pages from the web sites for which permission to distribute was granted. Some effort was made to eliminate encoded binary data from within news items but one VLC participant has indicated that this was not totally successful. Efforts were also made to eliminate web pages which explicitly claimed copyright for an organisation other than the host site.

Data within the `tar` files on the VLC tapes was formatted in the same way as the data on the CD-ROMS - as a directory hierarchy of multi-document files compressed using the standard Unix `compress` utility. Document identifiers were structured to allow unambiguous identification of collection, sub-directory and filename. Every document contained the four 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. No effort was made to ensure that resulting documents conformed to SGML standards.

5 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-6 AdHoc topics (301-350) over both the baseline and the VLC 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) for the VLC run compared with the baseline run.

All retrieved documents were judged. Only one baseline and one VLC run were permitted due to assessment resource limitations.

Participants were given the choice of comparing the measures for FIXED QUERIES derived either manually, interactively (e.g. over CD4 and CD5 in the AdHoc task) or automatically OR for queries which were expanded automatically over the dataset in use. No interaction with queries was permitted using either the baseline or the VLC collections.

6 The Measures

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

M2. Precision@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. (Modified to incorporate the size of the data set.)

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 CD-ROMs and DATs was excluded. The starting point was the compressed data files on disk after copying the CD-ROMs and unpacking the DAT tarfiles. M4 included the time to build all structures (such as inverted files) which are necessary to process the final query. Groups building phrase dictionaries, thesauri, co-occurrence matrices etc. for use in query building (NOT in query processing) were encouraged to report these times separately as M4R.

7 The Assessments

Three judges were employed to assess the VLC document pool. One was a PhD student and former research assistant in Asian Studies, another was a research assistant in Sociology and the other a recent Honours graduate in Economic History. The first judge was also employed in the TREC-5 pre-track. Some overlap between judges was organised as a sanity check and no significant discrepancies were found.

The document pool (derived from both baseline and VLC submissions) contained 8511 documents of which 2909 documents were judged relevant.

Of the total VLC pool, 1465 documents (17%) were also judged (against the same topic) by the NIST assessors as part of the AdHoc pool. NIST and ACSys judges agreed on 83% of cases.

8 Makeup of VLC Judgment Pool and Relevant Set

It would have been unfortunate had all of the documents in the VLC judging pool (or the VLC relevant set) come from CDs 4 & 5 or indeed from only the NIST-collected documents. Table 1 shows that this was not the case. As might be expected, given that the topics were not oriented toward the VLC data, the probability of a given document being selected by a retrieval system was significantly lower for the ACSys-collected documents than for the NIST-collected ones. Table 2 shows that USENET news documents were 6.7 times less likely to be retrieved than NIST-collected ones. The corresponding figure for ACSys-collected non-USENET documents was 3.0.

The probability that a document in the judging pool was relevant did not differ much between the NIST-collected and ACSys-collected, non-USENET documents. However, a USENET document in the pool was only 76% as likely to be judged relevant as other documents in the pool.

Table 3 shows the breakdown of ACSys-collected documents by individual collection. Perhaps surprisingly given the nature of some of the collections, each collection contributed at least one document to the relevant set.

8.1 Was the Baseline Collection an Unbiased Sample?

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 subset has been described above. The baseline subset contains 10.02% of the VLC data and 10.05% of the documents.

Of the 4833 different documents retrieved in the runs over the full VLC 460 (9.52%) were actually baseline documents. The proportion of documents in the VLC and the sample which were retrieved by VLC (not baseline) runs were 0.0006451 and 0.0006108. A test of one-sample

Table 4: Groups completing the VLC task. All groups attempted the full 20 gigabyte task but, due to problems, IBMg(Brown) actually used only 17.8 gigabytes.

Group	Query Gen.	Terms/Query	Stems	Query Opt.	Baseline Hardware	VLC Hardware
ANU	Auto_long	30	Yes	Yes	1 x DEC Alpha	8 x DEC Alpha
ATT	Auto_long	27	Yes	No	1 x SGI R10000	5 x SGI R10000
City	Auto_long	25	Yes	No	1 x Sun Ultra	1 x Sun Ultra
IBMs(Franz)	Auto_short	18 + Expand	Morphing	No	1 x IBM RS/6000	1 x IBM RS/6000
IBMg(Brown)	Auto_short	20	Morphing	No	1 x IBM RS/6000	6 x IBM RS/6000
UMass	Auto (title + desc)	66	Yes	No	1 x Sun Ultra	1 x Sun Ultra
U Waterloo	Manual	5.5	No	No	4 x Cyrix PC	4 x Cyrix PC

Table 5: M2: Precision at 20 documents retrieved. The asterisked items for IBMg(Brown) may have been higher if the full data had been used.

Group	Baseline	VLC	Ratio
City	0.320	0.515	1.61
ATT	0.348	0.530	1.52
ANU	0.356	0.509	1.43
UMass	0.387	0.505	1.31
IBMg(Brown)	0.275	0.361*	1.31*
U Waterloo	0.498	0.643	1.29
IBMs(Franz)	0.271	0.348	1.28

Table 6: M3: Average Query Processing Time (Elapsed minutes per 50 queries.) Figures in parentheses for IBMg(Brown) are scaled up by 20.1/17.8 to compensate for the smaller data size used. The baseline figure for the starred IBMs(Franz) run was derived by linear scaling of the VLC run.

Group	Baseline	VLC	Ratio
IBMg(Brown)	16.5	47.2(53.3)	2.86(3.23)
ANU	10.1	42.1	4.17
ATT	0.45	1.93	4.30
U Waterloo	0.189	1.12	5.93
City	6.3	61.4	9.75
IBMs(Franz)	886*	8857	10.0*
UMass	34.6	346	10.0

Table 7: M4: Data Structure Building Time (Elapsed Hours). Figures in parentheses for IBMg(Brown) are scaled up by 20.1/17.8 to compensate for the smaller data size used. The baseline figure for the starred IBMs(Franz) run was derived by linear scaling of the VLC run.

Group	Baseline	VLC	Ratio
ATT	0.768	2.57	3.34
IBMg(Brown)	3.23	28.4(32.1)	8.79(9.93)
IBMs(Franz)	86.9*	869	10.0*
UMass	6.85	69.14	10.1
City	9.9	103	10.4
U Waterloo	0.42	4.48	10.7
ANU	1.41	15.6	11.1

Table 8: MS: Data Structure Sizes (gigabytes). Figures in parentheses for IBMg(Brown) are scaled up by 20.1/17.8 to compensate for the smaller data size used. The baseline figure for the starred IBMs(Franz) run was derived by linear scaling of the VLC run. (Waterloo indicated at the conference that the sizes given in their questionnaire response and reported here may be higher than the correct values. Revised values are not yet available.)

Group	Baseline	VLC	Ratio
U Waterloo	3.36	30.9	9.20
City	2.47	23.6	9.55
ANU	0.626	6.06	9.68
IBMs(Franz)	1.21*	12.1	10.0*
IBMg(Brown)	1.21	10.8(12.2)	8.93(10.1)
ATT	1.23	13.02	10.6
UMass	1.22	11.43	?

Table 9: M5: Gigabyte-queries per hour per kilodollar.

	Baseline			VLC		
Group	Queries/Hr	kilo$	gB-Q/Hr/kilo$	Queries/Hr	kilo$	gB-Q/Hr/kilo$
U Waterloo	15873	7.44	4267.0	2678	7.44	7198.0
UMass	4392	45.7	3.8	439	45.7	3.8
ATT	6667	115	116	1554	394	78.9
City	476	14.2	67.0	48.9	14.2	68.8
ANU	297	23.9	24.8	71.3	95.1	15.0
IBMg(Brown)	182	17.3	21.0	63.6	123	10.3
IBMs(Franz)	3.39	30	0.226	0.339	30	0.226

proportion (with finite sample correction) shows that the sample proportion lies within the 95% confidence interval. Hence, there is no reason to conclude that the sample is biased with respect to proportion of retrieved documents.

9 Characteristics of Submitted Runs

The seven groups which passed the finishing post are listed in Table 4, which gives salient features of the methods used.

9.1 Hardware Used

A large range of hardware platforms were used, ranging from single workstations through clusters of PCs to large scale SMP systems. IBM, DEC, Sun, SGI and Cyrix hardware was used.

City used a single Sun workstation. UMass and ATT used a part of shared-memory multi-processor (SMP) systems. ANU, IBMs(Franz), IBMg(Brown) and Waterloo used networks or clusters of workstations (COWs).

Attempts to calculate "bang per buck" measures are not especially meaningful because:

1. 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 group which used a fraction of a very expensive system.

9.2 Approaches Taken

IBMg(Brown), ANU and ATT attempted to reduce the growth in query-processing time due to increased data size by adding more hardware. IBMs(Franz) actually did something similar but added all the individual times to produce a single-system time.

IBMg(Brown) used a collection fusion approach with no attempt to normalise rankings between the six parts of the collection. ATT divided the collection into 5 separately indexed pieces. Once indexes were built there was an exchange of document frequencies until all processors held correct global *df*s. ANU divided the collection and communicated *df* information (if necessary) at query processing time.

Waterloo used the same cluster of four PCs in both baseline and VLC runs. Waterloo also divided the collection into pieces but, due to use of distance-based relevance scoring, there was no resulting difference in results.

UMass and City essentially processed the VLC using a single processor although in the former case, the processor was one of four in an SMP system.

IBMs(Franz) was the only group not to run queries sequentially.

9.3 Query Generation

All query processing times reported were for the processing of fixed queries ie. did not include automatic feedback. City used automatic feedback over the collections but the query expansion time was not included in the tabulated figures.

Waterloo were the only group to use manually generated queries. These were the result of refinement by interaction with CD4/CD5 and other non-VLC documents. Other groups used automatic queries generated from all or various parts of the topic statement.

10 The Results

1. The shortest queries (5.5 terms, Waterloo) led to both the fastest processing and the best early precision. (Tables 4, 5 and 6) These queries were manually generated.

2. All runs showed at least 28% improvement in early precision for the VLC over the baseline. (Table 5)

3. Query processing time increased linearly with data size for uni-processor systems. Query processing time did not increase linearly for the Waterloo submissions which used the same hardware for both runs. (Table 6) It is understood that this is because a constant-time component of their algorithm ceased to be negligible when the data-size dependent component became very small, as was the case for their baseline run.

4. It is possible to reduce the query processing time scaling factor by scaling the hardware, but this year no group achieved a scaling factor of anything close to unity. (Table 6)

5. Data structure building is normally considered to be embarassingly parallel provided that the separately indexed pieces are evenly sized and not too small. However, only ATT exploited parallelism to bring the ratio below 10. (Table 7)

6. The fastest indexing rate was 7.84 gigabytes per elapsed hour (ATT) albeit on a very large machine. (Table 7)

7. Data structure sizes tended to increase linearly with the size of the raw data. (Table 8)

8. Data structure sizes for the VLC ranged from 6.06 gigabytes (ANU) to 30.9 gigabytes (Waterloo)[1]

9. Given the difficulties outlined above of assigning comparable dollar values to hardware actually used, it is difficult to place much emphasis on the results presented in table 9.

11 Discussion and Conclusions

The VLC track results clearly demonstrate that there are a number of retrieval systems for which query processing over 20 gigabytes is not at all daunting.

Good performance on the VLC size does not demand the use of exotic and expensive hardware. The best evidence for this conclusion is the Waterloo run over the full 20 gigabyte collection using an etherneted cluster of four commodity PCs whose total cost was only US$7,440. This run (using manually generated queries):

- retrieved an average 12.8 relevant documents in the first 20,
- indexed the data at a rate of 4.5 gigabytes per elapsed hour, and
- processed queries at a rate of 2678 queries per elapsed hour.

[1]This figure may not be correct. See the note in Table 8.

The only apparent downsides to the method used were the amount of disk space required and the use of manual queries.

The increase in early precision with the increase in data size is an interesting effect whose explanation is to be addressed elsewhere. It may be possible for groups to exploit the effect by using quicker, lower-quality algorithms on the VLC compared to the baseline. If judged well, early precision would remain constant but scalability would improve.

The processing of the 20 gigabyte collection should not be seen as an end in itself but rather as a way of predicting how retrieval systems will perform as data sizes grow to the multi-terabyte level. To make such predictions one must consider both the query processing performance of the system at a particular level and its scaling factor (after convincing oneself that the system will continue to scale at that rate). If query processing time grows in proportion to data size, then seconds at the gigabyte level will become hours at the multi-terabyte level. On the other hand, query processing times which remain constant despite data growth are not attractive if they already take hours at the gigabyte level.

12 The Future

12.1 Increasing Collection Size?

It is doubtful that increasing the size of the VLC by a small factor would improve the value of the track. Growth by a further order of magnitude would extend the scope of the problem but could dramatically increase the cost of participating in the track and possibly the cost of organising the track.

Considerable difficulty has been experienced in persuading organisations to make data available, due to concerns about data security or because of the resources required by the data donor to extract the data in a suitable form, or because the data holder itself does not have permission to distribute some of it.

Consequently, the only visible options for large increases in collection size are:

1. adding huge amounts of USENET news archived by the University of Waterloo;
2. approaching the Internet Archive (`http://www.archive.org/`);
3. replicating the existing data.

Participants in the VLC workshop at TREC-6 strongly expressed the view that an effort should be made to build the VLC up to 100 gigabytes for TREC-7, even if all the additional data is USENET news items. Interest was also expressed in addressing the problem of dealing with duplicate or near-duplicate documents.

12.2 Standardising Systems

It has been suggested that an attempt should be made to negate differences in hardware by defining a benchmark whose results could be used to scale timing results on the tasks. Unfortunately, it is likely that this would raise as many questions as it answered, as the algorithms employed differ enormously in the relative demands they place on CPU, memory, disk and network components.

12.3 Possible Revisions to Track Guidelines

It may be possible to allow more than one submission per group in 1998, provided that the size of the assessment pool does not grow too much.

12.4 Goals, Challenges and Purposes

The VLC track serves a number of different purposes:

- It complements mainstream TREC by allowing qualification, measurement and comparison of systems on the efficiency dimension.
- It may stimulate the development of algorithms whose space and time cost grows less rapidly than the increase in data size.
- It encourages consideration of the most suitable hardware and software architectures for tackling huge text collections of the (near) future.

This year failed to produce a set of results showing all of: query processing time over twenty gigabytes < 1.0 sec, precision@20 > 0.5, indexing rates > 10 gigabytes/hr and scaling factors ≈ 1.0. However, there are indications that such a combination may be possible and that achieving it may not require excessively expensive hardware.

Acknowledgements

We are indebted to Mark Sanderson and Jon Ritchie of Glasgow University who arranged for the release of data from the Glasgow Herald and for the additional data from the Financial Times and to Gordon Cormack and Rob Good of the University of Waterloo who supplied huge quantities of archived USENET news. Thanks also to Donna Harman for official support on behalf of NIST and the TREC Program Committee without which the VLC would not have come into being. Donna Harman, Ellen Voorhees and Dawn Tice(Hoffman) of NIST provided advice, support and practical assistance.

We acknowledge the work of Jason Haines, Tim Potter and Nick Craswell in formatting the new data and to Deborah Johnson, Sonya Welykyj and Josh Gordon in assessing submissions.

We would also like to express our appreciation to the organisations who gave permission to use their data in the VLC track. The willingness of organisations to make available commercially valuable data is a vote of confidence in TREC and in the integrity of its participants. VLC donor organisations in 1997 were: Canadian House of Commons (for Canadian Hansard); Australian Department of Defence (for the Defence Home Page); Australian Computer Society (for ACS web pages); National Library of Australia (for NLA web pages); Australian Broadcasting Commission (for Radio National web pages); Commonwealth Scientific and Industrial Research Organisation (for CSIRO web pages); Australian National University (for web pages); Victorian University of Technology (for web pages); Latrobe University (for web pages); Ballarat University (for web pages); Adelaide University (for web pages); Charles Sturt University (for web pages); University of Tasmania (for web pages); Edith Cowan University (for web pages); Murdoch University (for web pages); University of Newcastle, NSW (for web pages); Financial Times, London (for newspaper data 1988-1990); Caledonian Newspapers Ltd, Scottish Media Group (for Glasgow Herald data, 1995-97); Parliament of Australia (for parliamentary data including Hansard 1970-1995); CAUT Clearinghouse in Engineering (for web pages); Australian Attorney-General's Department (for legislation, court decisions and other legal data); Uniserve Coordinating Centre (for web pages); Australian Department of Industrial Relations (for industrial relations data).

Bibliography

HARMAN, D. K. Ed. 1992. *Proc. 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., THISTLEWAITE, P., AND CRASWELL, N. 1997. *TREC Very Large Collection (VLC) web page.* `http://pastime.anu.edu.au/TAR/vlc.html/`: ACSys Cooperative Research Centre, Australian National University.