Information Retrieval, 1, 115–137 (1999)

Scaling Up the TREC Collection*

DAVID HAWKING david.hawking@cmis.csiro.au
PAUL THISTLEWAITE paul.thistlewaite@cs.anu.edu.au
Cooperative Research Centre For Advanced Computational Systems
Department Of Computer Science, Australian National University, Canberra ACT 0200 Australia

DONNA HARMAN donna.harman@nist.gov
National Institute of Standards and Technology
Gaithersburg MD 20899

Received January 27, 1998; Revised October 14, 1998; Accepted October 14, 1998

Abstract. Due to the popularity of Web search engines, a large proportion of real text retrieval queries are now processed over collections measured in tens or hundreds of gigabytes. A new Very Large test Collection (VLC) has been created to support qualification, measurement and comparison of systems operating at this level and to permit the study of the properties of very large collections. The VLC is an extension of the well-known TREC collection and has been distributed under the same conditions. A simple set of efficiency and effectiveness measures have been defined to encourage comparability of reporting. The 20 gigabyte first-edition of the VLC and a representative 10% sample have been used in a special interest track of the 1997 Text Retrieval Conference (TREC-6). The unaffordable cost of obtaining complete relevance assessments over collections of this scale is avoided by concentrating on early precision and relying on the core TREC collection to support detailed effectiveness studies. Results obtained by TREC-6 VLC track participants are presented here. All groups observed a significant increase in early precision as collection size increased. Explanatory hypotheses are advanced for future empirical testing. A 100 gigabyte second edition VLC (VLC2) has recently been compiled and distributed for use in TREC-7 in 1998.

Keywords: test collection, very large databases, text retrieval

1. Introduction

The Text Retrieval Conference (TREC) series [19] is a Defense Advanced Research Projects Agency (DARPA) and National Institute of Standards (NIST) initiative, "to encourage research in information retrieval from large text collections."

Organizations represented at the TREC conference have traditionally been required to perform, in advance, an *ad hoc* retrieval task, structured as described in the next paragraph. In the task description, *topic* means an English-language statement of a researcher's information need, expressed in a form which might be given to a human research assistant or librarian. By contrast, *query* means an instruction or set of instructions to the retrieval system. Some systems are capable of interpreting a topic directly as a query, but others require queries to be expressed in a special language. TREC-6 topics include three fields: a two or three word *title*, a sentence-length *description* of what the topic is about, and a

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

paragraph-length *narrative* providing necessary background and specifying more precisely the types of documents which are or are not relevant.

Each organization participating in TREC is supplied with a test collection and a set of topics. Participants must use one of a variety of permissible methods to convert each topic into a query for their retrieval system, run the queries over the supplied collection and send a ranked list of document identifiers to NIST. The top n (e.g. 100) documents from each submitted list are pooled and judged by human assessors. Documents, topics and judgments from past TRECs are available as training data.

The overview of the proceedings of TREC-1 [11], 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 resulting 2-gigabyte collection used in TREC-1 was two orders of magnitude larger than previous collections, and was thus 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 organizations, newspapers, businesses and government departments. However, it is clear that some organizations 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 this scale in the commercial world. For example, as of September 1998, the Lexis-Nexis online service was claimed [18] to include nearly 4 terabytes (1 terabye = 1024 gigabytes) of data, albeit divided into many separate collections.

In August 1998, the World Wide Web was believed to be nearly as large and growing rapidly. At that time, the Alexa Corporation claimed [1] that the amount of accessible data (i.e. reachable from some other page via a hyperlink and not subject to access controls) on the Web was approximately three terabytes and that this quantity was doubling every eight months. Some Web search engines now support searches over a large fraction of this data. For example, as at May 1998, Alta Vista indexes were stated to cover 140 million pages [8]. This corresponds to nearly one terabyte of text.

Given the huge volume of requests handled by such search engines[1], it is now the case that a high proportion of all text retrieval is actually carried out over data sets measured in tens or hundreds of gigabytes.

Accordingly, in line with the initial TREC charter of realism, an expanded test collection, known as the Very Large Corpus or VLC, has been created. The first edition, distributed in 1997, comprises 20.14 gigabytes of data and 7.49 million documents. The VLC includes all of the TREC adhoc data and is intended to be used as a supplement to mainstream TREC resources. It is available to participants under the terms of the standard TREC data permission agreement. Details of TREC data (issued on five CD-ROMs) and data permission agreements are available via the TREC web page. [19] A second edition VLC of 100 gigabytes (known as the VLC2) has recently been compiled and distributed. [15]

The first edition of the VLC was used in a special interest track within the 1997 TREC conference (TREC-6). [25] Fourteen groups received data tapes and seven submitted runs. The VLC2 corpus is currently being used in TREC-7.

The main emphasis of the TREC conference has been on the *effectiveness* of retrieval systems as measured by *precision* (the proportion of retrieved documents which are in fact relevant) and *recall* (the proportion of all relevant documents which have been retrieved). There is evidence to suggest that the effectiveness of text retrieval systems has improved significantly since 1992 [5]. However, it has not been clear whether algorithms which work well on the TREC task would operate successfully over much larger collections. The availability of the VLC and the definition of a set of standardised scalability metrics should allow this question to be answered.

By contrast, Web search engines are forced by economic considerations and the huge query-processing demand to emphasise the *speed* of query processing. Until now it has been difficult to measure their effectiveness because the document collection represented by the Web is quite dynamic and because each search engine indexes a different subset of the available data. It is hoped that a well-defined static test collection of similar size will overcome these problems.

The value of a test collection lies not only in the data itself but also in the availability of relevance judgments for a large set of research topics. Complete sets of judgments are available for small test collections but are not affordable for larger ones. If a judge can read one megabyte (approximately the amount of text in a typical book) of data per day or 250 megabytes per year, then one gigabyte takes four years and one hundred gigabytes takes four centuries *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. In the overview of TREC-4 [12, p. 7, Table 4] it is reported that the average number of judged documents per topic in the TREC-4 adhoc task was 957 out of a collection of over 500,000 documents and that the corresponding figures for the first three TRECs varied from 703 to 1466 out of a collection of over 700,000 documents. The same paper presents results of completeness tests applied to the TREC-2 and TREC-3 pools and concludes that the judgments are "acceptably complete" for the effectiveness measures used in TREC.

Assessment resources available to support the VLC are not sufficient to support measures that depend on estimating recall over a 20-gigabyte (or larger) collection. Even if sufficient resources were available to support the TREC pooling method, that method is not likely to be as effective in the VLC context. For any given topic there may be ten (or fifty, in the case of the second edition) 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 defined for the VLC track are confined to *early precision*, i.e. precision achieved at a small, fixed number of documents retrieved. It is envisaged that retrieval system developers will characterise the effectiveness of their system on the main TREC AdHoc task and then observe how speed and effectiveness scales with increasing collection size.

A consequence of incomplete assessments is that the judgments which are made have little value for assessing further searches. However, the assessment resources required to support early precision measures are sufficiently modest to make self or mutual assessment feasible for many groups.[2]

This paper outlines the goals of the VLC effort, analyses and reports the composition of the first edition, defines a set of measures, reports results obtained in the TREC-6 track, discusses some theoretical questions which arise and presents conclusions and suggestions for further work.

2. Goals of the Very Large Collection

A cornerstone of TREC philosophy is the encouragement of a diversity of experimentation. A minimal set of guidelines is issued to increase the comparability of results but participants are free to conduct whatever experiments they choose using the supplied data.

Within this flexible framework, the following lists illustrate the range of questions which it is hoped may be investigated with the aid of the VLC.

1. System qualification, measurement and improvement:
   - (A) Will a particular system work over a given collection size? At what point does it cease working properly, due to design, hardware or operating system limits?
   - (B) How fast will it process queries over a given collection size?
   - (C) How does its query processing speed decay as data grows?
   - (D) What is the effect of a particular algorithmic improvement?
   - (E) Does one particular system work better than another?
2. Investigation of the efficiency dimension in Text Retrieval:
   - (A) What is the relationship between collection size and scale of hardware required?
   - (B) Searchers need both retrieval effectiveness and query processing speed. Is there necessarily a tradeoff between them?
   - (C) Which retrieval techniques work best over large collection sizes?
   - (D) Do some algorithms scale better than others?
   - (E) What optimization techniques are most effective in reducing query processing time in the context of very large collections? See [4, 21] for examples of optimization techniques.
3. Architectures for future digital libraries:
   - (A) What sort of hardware is needed to run queries over collections whose size is in excess of one hundred gigabytes?
   - (B) What architecture is most cost effective for a given combination of collection and query processing load?
   - (C) Can scalable hardware (clusters of workstations) be used to support scalable retrieval software?
   - (D) Should large collections be processed as single units or are there advantages to using collection selection and result merging techniques?

4. Investigation of special features of large collections:

   (A) How should the increased problem of similar or duplicate documents be handled?

   (B) Are there special problems due to increased diversity of source and format?

   (C) Is it easier to find relevant documents in large collections?

As may be seen, most of the questions listed above relate to the engineering of text retrieval systems. One of the more theoretical issues ("Is it easier to find relevant documents in large collections?") is addressed in Section 6, and in a follow-up paper.

The VLC may provide a realistic test environment for research into distributed information retrieval problems such as server selection and result merging, provided that necessary relevance judgments can be made. It is also likely that studies will be based on particular subsets of the data, such as Web data, NEWS data, legal data, parliamentary data, or concurrent newspaper data.

3. The Data

With support from NIST and the TREC program committee, the first edition of the VLC was assembled and distributed by the Advanced Computational Systems Cooperative Research Centre (ACSys), whose core participants are the Australian National University, the Commonwealth Scientific and Industrial Research Organization, Fujitsu, Sun and DEC. ACSys also organised the TREC-6 VLC track and employed the VLC assessors. Additional information on the Very Large Collection is available on the VLC Web page [15].

The 20.14 gigabyte first edition collection (including all data from TREC CD-ROMs 1-5, referred to for convenience as "NIST" data) was assembled with assistance from a large number of data holders. The additional ("ACSys") data was distributed on DAT (4mm) tapes due to logistical and financial difficulties with using CD-ROMs.

3.1. The Need for a Baseline Sample

At least as important as measures taken over the VLC, are the ratios of those measures from one collection size to another. Of particular interest in scalability experiments are order-of-magnitude steps in collection size. Ideally, the smaller collection should exhibit the same properties (scaled down) as the larger.

Accordingly, from the 20 gigabyte first edition VLC, a uniform 10% sample was defined for use as a baseline. 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 length dropped to 2.8 MB from 6.2 MB. The 10% sample was distributed on a separate tape.

3.2. Composition of the First Edition VLC Data

The sub-collections of the first edition VLC are listed in Table 1. As may be seen, the ACSys data is somewhat biased by the inclusion of roughly 8.7 gigabytes of USENET news postings

Table 1. Sub-collections of the VLC. NIST data comprises the five CDs, while ACSys data comprises the four DATs.

Sub-coll.	Source	# Doc.s	Smallest Doc. (kB)	Largest Doc. (kB)	Size (MB)
CD1 AP	Associated Press	84,678	0.19	16.1	254.2
CD1 DOE	U.S. Energy Dept.	226,087	0.06	2.1	183.8
CD1 FR	U.S. Federal Register	25,960	0.09	2575.5	259.7
CD1 WSJ	Wall Street Journal	98,732	0.07	78.6	266.6
CD1 ZIFF	Ziff-Davis Pub.s	75,180	0.30	359.1	242.3
	CD1 total	**510,637**			**1.18 gB**
CD2 AP	Associated Press	79,919	0.18	15.0	237.2
CD2 FR	U.S. Federal Register	19,860	0.09	1,835.3	209.2
CD2 WSJ	Wall Street Journal	74,520	0.14	132.9	241.9
CD2 ZIFF	Ziff-Davis Pub.s	56,920	0.34	153.8	175.5
	CD2 total	**231,219**			**0.84 gB**
CD3 AP	Associated Press	78,321	0.17	16.6	237.5
CD3 PATENTS	U.S. Patents	6,711	1.25	609.8	242.6
CD3 SJM	San Jose Mercury	90,257	0.60	61.6	286.8
CD3 ZIFF	Ziff-Davis Pub.s	161,021	0.28	500.7	344.8
	CD3 total	**336,310**			**1.09 gB**
CD4 CR	Congressional Rec.	27,922	0.23	3,860.4	235.4
CD4 FR94	U.S. Federal Register	55,630	0.10	6,239.4	394.9
CD4 FT	Financial Times	210,158	0.28	185.7	564.1
	CD4 total	**293,710**			**1.17 gB**
CD5 FBIS	Foreign Broadcast	130,471	0.50	914.8	470.2
CD5 LATIMES	LA Times	131,896	0.27	938.6	475.3
	CD5 total	**262,367**			**0.92 gB**
DAT1 NEWS01	USENET News	446,106	0.17	418.3	954.5
DAT1 NEWS02	USENET News	450,027	0.24	1,165.4	943.1
DAT1 NEWS03	USENET News	482,395	0.17	966.0	936.6
DAT1 NEWS04	USENET News	483,145	0.14	475.0	966.0
	DAT1 total	**1,861,673**			**3.71 gB**
DAT2 AAG	Attorney General's	561,566	0.07	1,925.0	1874.5
DAT2 ADIR	Industrial Relations	42,841	0.12	4,757.2	775.0
DAT2 NEWS05	USENET News	590,202	0.09	691.9	1169.7
DAT2 PGUT	Project Gutenberg	3,303	0.07	5,016.5	430.3
	DAT2 total	**1,197,912**			**4.15 gB**
DAT3 APLT	Aust. Parliament	421,681	0.21	697.7	1539.8
DAT3 AUNI	Uni. Web Sites	81,334	0.05	4,579.2	724.8
DAT3 GH	Glasgow Herald	135,477	0.18	135.8	393.6
DAT3 NEWS06	USENET News	571,891	0.14	493.1	1120.6
	DAT3 total	**1,210,383**			**3.69 gB**
DAT4 FT	Financial Times	202,433	0.26	58.1	526.7
DAT4 NEWS07	USENET News	520,282	0.21	605.8	1080.1
DAT4 NEWS08	USENET News	856,609	0.15	919.7	1727.9
DAT4 WEB01	Misc. Web pages	8,513	0.06	590.4	141.9
	DAT4 total	**1,587,837**			**3.40 gB**
Grand Totals		**7,492,048**			**20.14 gB**

to make up the target 20 gigabytes. This data has a significantly different character from the data on CDs 1-5. However, the remainder of the ACSys data 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 (part of WEB01 sub-collection) and Australian Hansards) and newspapers (eg. Glasgow

Herald and Financial Times). For the first time in TREC, HTML documents downloaded from the Internet are included (eg. from the Australian Broadcasting Commission, National Library of Australia and various university websites). Also for the first time in TREC, there is a large quantity of legal data, including both laws and judgments, made available by the Australian Attorney General's Department (AAG). The latter is mostly in HTML format.

Sub-collections of the first edition ACSys data are typically larger than those on CDs 1-5. For example, the AAG sub-collection totals 1874.5 megabytes whereas the largest NIST sub-collection totals 564.1 megabytes (FT sub-collection on CD4). However, the addition of the new data has not altered the minimum or maximum document length figures. Average document length has declined slightly, from 3.2 kilobytes for CDs 1-5 to 2.8 kilobytes for the entire VLC first edition.

3.3. Restrictions on Access to the VLC Data

Access to the VLC data is restricted to TREC participants and is subject to the terms and conditions of the TREC data permission forms. [19] Data was obtained from copyright owners on this basis. Copyright owners for the ACSys data are listed in the Acknowledgements below. Permissions were obtained from controllers of all websites used as sources of documents.

3.4. International Balance

The international balance of the data is significantly different from 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% 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 believed to be negligible.

3.5. Formatting

A variety of `flex` programs [20] and `perl` scripts [27] were used to convert supplied data into VLC format. The `wget` program [10] 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-track participant has indicated that this was not totally successful. Efforts were also made to eliminate web pages which explicitly claimed copyright for an organization other than the host site.

Data within the archive (Unix `tar`) files on the VLC tapes was formatted in the same way as the data on the CD-ROMS. Each sub-collection is represented by a directory hierarchy of multi-document files compressed using the standard Unix `compress` utility. Document identifiers are structured to allow unambiguous identification of collection, sub-directory and filename. Every document contains the four essential "SGML" markers delimiting documents and document identifiers. A program `coll_check`[3] was used to check that

Table 2. Crude breakdown of first edition VLC, TREC-6 VLC assessment pool and relevant set by source. NIST documents refer to the TREC documents distributed by NIST on five CD-ROMs. (Hawking et al., 1997b) TREC-6 documents are those contained on the fourth and fifth NIST CD-ROMs. ACSys documents refer to the additional documents distributed by ACSys. The number of documents judged for a topic (the size of the assessment pool) is the cardinality of the union of the sets of documents submitted.

Source	# Documents	# Documents judged	# Relevant documents
Total VLC	7,492,048(100%)	8511(100%)	2909(100%)
All NIST doc.s	1,634,243(21.8%)	5034(59.2%)	1833(63.0%)
- TREC-6 doc.s	556,077(7.4%)	1608(18.9%)	631(21.7%)
- Other NIST doc.s	1,078,166(14.4%)	3426(40.3%)	1202(41.3%)
All ACSys doc.s	5,857,805(78.2%)	3477(40.9%)	1076(37.0%)
- USENET news doc.s	4,400,657(58.7%)	2001(23.5%)	552(19.0%)
- ACSys non-USENET doc.s	1,457,148(19.4%)	1476(17.3%)	524(18.0%)

Table 3. TREC-6 VLC. Proportion of documents retrieved and proportion of documents judged to be relevant for different sources. (Obtained by dividing the raw frequencies in Table 2 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	Prop. Ret.	Prop. Rel.	Pr(relevant|retrieved)
All NIST doc.s	3.08×10^{-3}	1.12×10^{-3}	0.364
- TREC-6 docs.	2.89×10^{-3}	1.13×10^{-3}	0.392
- Other NIST docs.	3.18×10^{-3}	1.11×10^{-3}	0.351
All ACSys docs.	5.93×10^{-4}	1.84×10^{-4}	0.309
- USENET news docs.	4.55×10^{-4}	1.25×10^{-4}	0.276
- ACSys non-USENET docs.	1.01×10^{-3}	3.60×10^{-4}	0.355

each document conforms to this elementary structure and that document identifiers were unique. No effort has been made to ensure that resulting documents conform to SGML standards. Indeed, there is a plethora of unmatched angle-brackets in the NEWS collections.

4. Measures Over The VLC

The following measures have been defined for use in conjunction with the VLC. It is assumed that the retrieval system being tested will index the full VLC and process a set of 50 queries over the indexes thus built. For each query, the system must identify and rank the 20 documents best matching the topic from which the query was derived.

M1. Task Completion. (Is the system actually capable of indexing the VLC data set and processing queries over it?)

M2. Precision@20. (Each document in each ranked list is judged by a human assessor for relevance to the topic. The measure is the proportion of the top 20 documents which are judged to be relevant.)

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. (The query processing rate is multiplied by the size of the collection and divided by the cost of the hardware. This is an attempt to measure the cost-normalised rate at which query processing work is being performed. It is assumed that the amount of work being done is proportional to the size of the collection. The kilodollar figure used is an estimate of the current list price in U.S. dollars of the aggregation of system hardware components actually used in performing the query processing task.)

MS. Size of all data structures needed for query processing. (Gigabytes, not including compressed or uncompressed raw data.)

All the timing measures are based on elapsed time rather than CPU time.

M4 represents the minimum possible elapsed time from receiving the data until the data structures necessary to process the queries used in M3 are built, using the chosen hardware and indexing software. Time to actually read the media on which the compressed raw data is distributed is not included. The starting point is the compressed data files on disk after copying the CD-ROMs and extracting all the files from tape. M4 thus includes the time to build all structures (such as inverted files) which are necessary to process the final queries.

It is expected that the way the measures scale with increasing collection size will be as interesting or more interesting than the absolute values at a particular collection size.

5. Use of the VLC in TREC-6

Participants in the TREC-6 VLC track were required to process queries generated from the TREC-6 AdHoc topics (301-350) [19] over both the baseline and the VLC datasets and to return for assessment the first 20 documents retrieved in each case. In this section and the next, the term VLC will be used to refer to the first edition VLC and the term baseline will refer to the standard 10% sample.

Elapsed times 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 measures defined in Section 4 above for the VLC run compared with the baseline run. Full guidelines are available from the VLC web page [15].

5.1. Assessments

Three judges were employed to assess the TREC-6 VLC document pool. One was a PhD student and former research assistant in Asian Studies, another was a research assistant in Sociology and the third a recent Honours graduate in Economic History.

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.

Table 4. TREC-6 VLC. Contributions of individual ACSys collections to the pool and the relevant set. For comparison, the last line shows the corresponding data for the NIST documents. The *NIST ratios* represent an attempt to compare the usefulness (to the particular topic set on a per document basis) of the ACSys sources relative to that of the NIST documents. For each ACSys source, the proportion of documents which formed part of the pool (or part of the relevant set) was divided by the corresponding proportion for all NIST-distributed documents to give the NIST ratio. NIST ratios for both the PGUT and WEB01 sources exceed unity, indicating that a randomly chosen document from either of these collections would be more likely to be retrieved by the systems and more likely to be judged relevant than a similarly chosen NIST document.

	Collection	Pool			Relevant Set		
Source	# docs	# docs	% of pool	NIST Ratio	# docs	% of rel. set.	NIST Ratio
AAG	61,566	230	2.7%	0.133	59	2.0%	0.094
ADIR	42,841	9	0.1%	0.068	1	0.0%	0.021
APLT	421,681	501	5.9%	0.386	185	6.4%	0.391
AUNI	81,334	134	1.5%	0.535	40	1.4%	0.438
FT	202,433	259	3.0%	0.415	100	3.4%	0.440
GH	135,477	251	2.9%	0.601	107	3.7%	0.704
NEWS01	446,106	180	2.1%	0.131	65	2.2%	0.130
NEWS02	450,027	221	2.6%	0.159	63	2.2%	0.125
NEWS03	482,395	228	2.7%	0.153	56	1.9%	0.104
NEWS04	83,145	233	2.7%	0.157	61	2.1%	0.113
NEWS05	590,202	325	3.8%	0.179	91	3.1%	0.137
NEWS06	571,891	260	3.1%	0.148	47	1.6%	0.073
NEWS07	520,282	240	2.8%	0.150	60	2.1%	0.103
NEWS08	856,609	314	3.7%	0.119	109	3.7%	0.113
PGUT	3,303	30	0.4%	2.949	5	0.2%	1.350
WEB01	8,513	62	0.7%	2.364	27	0.9%	2.828
All NIST	1,634,243	5,034	59.2%	1.000	1,833	63.0%	1.000

5.2. Composition 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 documents. Table 2 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 documents than for the NIST ones. Table 3 shows that USENET news documents were 6.8 times ($3.08 \times 10^{-3}/4.55 \times 10^{-4}$, using the figures in column 2) less likely to be retrieved than those from NIST. The corresponding figure for ACSys non-USENET documents was 3.0 ($3.08 \times 10^{-3}/1.01 \times 10^{-3}$).

The probability that a document in the judging pool was relevant differed only slightly between the NIST and ACSys, 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 4 shows the contributions of individual ACSys sub-collections to the assessment pool and to the relevant set. Perhaps surprisingly, given the nature of some of the data, each sub-collection contributed at least one document to the relevant set. For example, the Project Gutenberg sub-collection contains the *CIA Factbook*, which was relevant to a number of topics!

Table 5. Basic characteristics of VLC runs submitted in TREC-6. *Query Gen.* indicates how queries were generated from the topic statements. The letters T (title), D (description) and N (narrative) indicate the topic statement fields which were used by automatic query generators. *Terms/Query* is a measure of query complexity and indicates the average number of literal strings (words, stems or phrases, but not query operators) in the queries employed. *Query Opt.* indicates whether or not any query optimization techniques were applied. All groups attempted the full 20 gigabyte task but, due to data-handling problems, IBMg actually used only 17.8 gigabytes.

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

5.3. 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 in Section 3.1. 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 proportion (with or without 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.

5.4. Characteristics of Submitted Runs

VLC submissions were received from the following groups: Australian National University / ACSys (ANU/ACSys - [14]); AT&T (ATT - [23]); City University, London (City - [26]); IBM T.J. Watson Research Center (IBMs - [9]); IBM T.J. Watson Research Center (IBMg - [3]); University of Massachusetts at Amherst (UMass - [2]); University of Waterloo (Waterloo - [6]). Salient features of these seven official submissions are tabulated in Table 5. Sections 5.5 to 5.6.6 discuss the questions addressed by these runs and the methods and equipment used.

5.5. Hardware Used

A large range of hardware platforms were used, ranging from single workstations through clusters of PCs to large scale parallel 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 Symmetric Multi-Processor (SMP) systems. ANU/ACSys, IBMs, IBMg and Waterloo used networks or clusters of workstations (NOWs or COWs).

Attempts to calculate "bang per buck" type measures, such as M5, or to compare algorithms independently of the hardware on which they run are not especially meaningful because:

1. Groups used available hardware rather than explicitly selecting it for the task. Their software may have run as fast or faster on much cheaper hardware.

2. Some groups were unable to run their systems in dedicated mode. In these cases, no attempt has been made to control for the effect of other users.

3. It is difficult to determine a meaningful dollar value for a fraction of a very expensive system or for superseded hardware.

5.6. Questions Addressed

Many of the questions listed under the VLC goals (Section 2) were addressed by individual groups or may be addressed by looking at the group of runs.

5.6.1. System Qualification (Goal 1) In the case of City and UMass, the main goal of VLC participation was to apply a standard retrieval system on a single processor[4] to the 20 gigabyte task in order to confirm correct operation and to validate expectations of a linear increase in processing times. Of the fourteen groups who received the VLC data, six may be seen to have achieved the basic qualification goal (i.e. demonstrating their system's ability to index 20 gigabytes of data and process a set of queries over it) within the context of the VLC track.

5.6.2. Reducing Resource Requirements (Goal 2) Optimization of query processing has the potential to reduce query processing time for a given collection size and also to produce a better-than-linear relationship between query processing and data size. In the TREC-6 track this potential was only minimally exploited. ANU/ACSys optimised queries both on the baseline and on the VLC by processing only the 15 terms with the lowest df values, where df is widely-used shorthand for the number of documents in the collection which contain the term of interest.

UMass reported a halving of processing time (for an unofficial run) with no loss in precision by updating the scores of only the top-scoring 1000 documents for each individual term.

Various techniques were used to reduce index size and thus to lower disk space and memory demands. For example, City reported that they removed the term-position information usually held in their indexes and other groups (such as ANU/ACSys) used compression techniques.

5.6.3. Scalability (Goals 2 and 3) IBMg, ATT and ANU/ACSys attempted to reduce the growth in query-processing time due to increased collection size by using scalable hardware[5]. The simplest example of scalable hardware is a cluster of workstations whose entry-level processing power can be increased by two or more orders of magnitude, in unit (workstation) increments. In other words, an organization may initially purchase a small number of workstations (possibly only one) and add workstations in response to growth in the amount of on-line data.

ANU/ACSys attempted to confirm an earlier result [13] showing that, with certain restrictions, if the number of workstations in a cluster is increased in proportion to the size of the data, then query processing time can be held constant.

When using a cluster of workstations, it is natural to divide the data into a number of separately indexed sub-collections, one or more per node. This introduces the well-known collection fusion problem. For systems which rank documents using weights based on the inverse of term df (the number of documents containing the query term), the problem is that if local df values are used, relevance scores across sub-collections are not generally comparable.

IBMg defined six isolated sub-collections of the VLC and made no attempt to normalise scores across them. Unlike most other groups, they derived queries from only the description field of the topic statements, rendering it impossible to measure the loss of precision, if any, resulting from this decision[6].

ATT divided the collection into five separately indexed pieces. Once local indexes were built, a once-only exchange of document frequencies replaced local dfs for all terms with correct global values. The exchange process involves collecting, summing, and broadcasting df values. In the ANU/ACSys case the collection was also divided across nodes but the df exchange occurred at query processing time and involved only the query terms. These two approaches should both yield correctly merged results but the ATT approach is probably more efficient as the small once-only exchange cost at build time would generally be smaller than the accumulation of smaller query-time costs in the ANU/ACSys case.

5.6.4. Cost-Effectiveness and Fault Tolerance (Goal 3) Waterloo also investigated the benefits to be obtained from scalable clusters of workstations. They identified low cost and potential fault tolerance as additional benefits of the approach. Commodity PCs were used as they were seen to be attractive to small-scale Internet information providers and because they would minimise the cost of large scale clusters.

Waterloo used the same cluster of four PCs in both baseline and VLC runs. Like the groups listed above, they divided the collection but, due to their use of distance-based relevance scoring (which does not rely on df information), [6, Section 2] there was no consequent difference in results.

5.6.5. Query Generation All query processing times reported were for the processing of fixed queries. City used *automatic feedback* over the collections but the query expansion time was not included in the reported figures. Automatic feedback is a process in which an initial query is used to select a number of highly ranked documents which are assumed to

Table 6. M2: Precision at 20 documents retrieved. TREC-6 VLC track. The asterisked items for IBMg 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/ACSys	0.356	0.509	1.43
UMass	0.387	0.505	1.31
IBMg	0.275	0.361*	1.31*
Waterloo	0.498	0.643	1.29
IBMs	0.271	0.348	1.28

be relevant and which are mined for terms to be included in the final query. Only the time to run the final query was reported.

Waterloo was 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 fields of the topic statement. ANU/ACSys, ATT and City used the full statement, and UMass used title plus description fields. Both IBM groups used only the description field. (Experience with the same topic statements in the TREC-6 Adhoc tasks suggests that use of description-only would lead to significantly reduced retrieval effectiveness.)

5.6.6. Batch vs. Timesharing of Queries In heavy query-load situations, new queries typically arrive at a search engine before processing of their predecessors is complete[7]. System implementors have a choice of processing queries as a sequential stream (batch mode) or processing multiple queries concurrently (timesharing mode). It is well known that, for simple systems, batch mode achieves greater throughput but that long queries in the batch may cause excessive delay to short ones. However, in more complex systems, it may be possible to exploit pipelining techniques, parallel hardware, or better memory referencing locality to transfer the advantage to the timesharing mode.

Only the IBMs group did not run queries sequentially. Unfortunately, insufficient information is available to compare this timesharing run with a batch alternative, either on throughput or on waiting times.

5.7. TREC-6 VLC Results

Tables 6 to 10 show the results achieved by participants in the TREC-6 VLC track. Cross-system comparisons must be made with considerable caution since participating groups differed markedly in their experimental goals, in the resources available to support their efforts, in the way they generated queries and in the degree to which the hardware used was shared with other users.

1. The shortest queries (5.5 terms, Waterloo) led to both the fastest processing and the best early precision (Tables 5, 6 and 7). These queries were manually generated.

Table 7. M3: Average Query Response Time (Elapsed seconds per query). TREC-6 VLC track. Figures in parentheses for IBMg are scaled up by 20.1/17.8 to compensate for the smaller collection size used. The baseline figure for the starred IBMs run was derived by linear scaling of the VLC run.

Group	Baseline	VLC	Ratio
IBMg	19.8	56.6(63.4)	2.86(3.23)
ANU/ACSys	12.1	50.5	4.17
ATT	0.54	2.32	4.30
Waterloo	0.23	1.34	5.93
City	7.6	73.7	9.75
IBMs	1063	10,628	10.0*
UMass	41	410	10.0

Table 8. M4: Data Structure Building Time (Elapsed Hours). TREC-6 VLC track. Figures in parentheses for IBMg are scaled up by 20.1/17.8 to compensate for the smaller collection size used. The baseline figure for the starred IBMs run was derived by linear scaling of the VLC run.

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

Table 9. MS: Data Structure Sizes (gigabytes). TREC-6 VLC track. Figures in parentheses for IBMg are scaled up by 20.1/17.8 to compensate for the smaller collection size used. The baseline figure for the starred IBMs run was derived by linear scaling of the VLC run. The sizes quoted for City include space needed during query expansion. (Query execution was reported as requiring 0.8 and 5 gigabytes for the baseline and VLC respectively.)

Group	Baseline	VLC	Ratio
UMass	1.22	11.43	9.37
Waterloo	1.7	16	9.4
City	2.47	23.6	9.55
ANU/ACSys	0.626	6.06	9.68
IBMs	1.21*	12.1	10.0*
IBMg	1.21	10.8(12.2)	8.93(10.1)
ATT	1.23	13.02	10.6

2. All runs showed at least 28% improvement in early precision for the VLC over the baseline (Table 6).

Table 10. M5: Gigabyte-queries per hour per kilodollar. TREC-6 VLC track. See Section 4 for an explanation of the measure. Dollar values are estimated 1997 list prices (in U.S. dollars) of the hardware actually used in query processing.

	Baseline			VLC		
Group	Queries/Hr	kilo$	gB-Q/Hr/kilo$	Queries/Hr	kilo$	gB-Q/Hr/kilo$
Waterloo	15,873	7.44	4267.0	2678	7.44	7198.0
ATT	6667	115	116	1554	394	78.9
City	476	14.2	67.0	48.9	14.2	68.8
ANU/ACSys	297	23.9	24.8	71.3	95.1	15.0
IBMg	182	17.3	21.0	63.6	123	10.3
UMass	87.8	45.7	3.84	8.78	45.7	3.84
IBMs	3.39	30	0.226	0.339	30	0.226

3. Query processing time increased linearly with collection 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 7). 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 no group achieved a scaling factor approaching unity (Table 7).

5. Data structure building is normally considered to be highly parallelizable, provided that the separately indexed pieces are evenly sized and not too small. However, only ATT exploited parallelism to bring the ratio significantly below 10 (Table 8).

6. The fastest indexing rate was 7.84 gigabytes per elapsed hour (ATT) albeit on a very large machine (Tables 8 and 5).

7. The data structure size ratios tended to approximate the ratios of the collection sizes. (Table 9).

8. Data structure sizes for the VLC ranged from 6.06 gigabytes (ANU/ACSys) to 23.6 gigabytes (City). The size quoted for City includes space needed during query expansion. (Query execution was reported as requiring 5 gigabytes for the VLC.)

9. The gigabyte-queries per hour per kilodollar measure (M5 in Section 4) seems to have some validity:

   (A) The scalability ratios in Table 7 for the two single-processor submissions (City and UMass) correspond almost exactly to the ratio of the two collection sizes, giving support to the idea that query processing work, in general, tends to be proportional to the size of the collection.

   (B) Earlier results [13] suggest that, using a cluster of workstations, query processing times can be held constant, provided that the number of workstations (and hence the system cost) is increased in proportion to the size of the collection.

   (C) Generalising from the two preceding sub-points to an ideally scalable system suggests that, for a given retrieval system and a given set of queries, the M5

measure should remain constant despite addition or removal of workstations and despite increases or decreases in the collection size. Differences in M5 scores for real systems using comparable queries could therefore be expected to highlight differences in algorithmic efficiency or departures from linearity.

Unfortunately, system cost is not a good measure of its query processing power, and (as noted above) it is difficult to assign comparable dollar values to hardware actually used. Consequently, it would be unwise to place much emphasis on the results presented in table 10.

6. Why does Precision@20 Increase?

The fact that all groups participating in the TREC-6 VLC track recorded a relatively large increase in precision@20 when moving from the sample to the full collection appears to be contrary to the expectation of early workers in the field of information retrieval. Salton and McGill [22] stated that precision can be expected to *decrease* when collection size increases because, "the number of retrieved and relevant documents [i.e. the number of retrieved documents which are actually relevant] is not likely to increase in proportion to the size of the collection[8]."

It is possible that Salton and McGill were assuming that *generality* (the probability that an arbitrarily chosen document in the collection is relevant to a query) would decline as the collection size increased. In fact, in the present case, generality should on average remain constant because the VLC baseline is a representative sample of the VLC. However, this would lead to a prediction of constant rather than increased precision for the larger collection.

The observed increase in precision is thus worthy of some analysis. Such an analysis may lead to a better understanding of the retrieval properties of random samples drawn from a large collection. It would be useful to be able to predict the increase in early precision resulting from scaling up a collection by a specified amount.

Clearly, there are more relevant documents in the superset collection but this fact does not, by itself, explain the observation. If the number of relevant documents increases by a factor of n, then so should the number of irrelevant documents which achieve high scores.

A number of (not necessarily mutually inconsistent) hypotheses have been informally suggested by TREC participants to at least partially explain the observed increase in precision@20:

Hypothesis 1: Baseline precision@20 measurements are necessarily lower because large numbers of topics have too few relevant documents in the baseline sample to achieve the precision@20 scores observed on the full VLC. For example, no system can possibly achieve baseline precision@20 higher than 0.1 on a topic for which there are only two relevant documents in the baseline collection. However, on the same topic, the full VLC may contain ten times as many relevant documents and precision@20 could reach unity for a perfect retrieval system.

Hypothesis 2: The first edition VLC is equivalent to a replicated collection in which each baseline document is repeated ten times. From this, it is expected that

the precision@20 on the 20 gigabyte collection should equal the precision@2 value for the baseline, when averaged over a sufficiently large number of topics.

Hypothesis 3: Precision-recall curves, as plotted for TREC, represent a kind of operating characteristic for the combination of retrieval system, query and spread of documents represented by the VLC. Because the baseline is a representative sample of the VLC, the operating characteristic is the same in both cases. If precision monotonically declines with increasing recall, as is usually the case, the probability that a particular document in the ranked list returned by a system is relevant also declines with increasing recall. The effect of increasing the collection size (adding both relevant and irrelevant documents) is to change the number of relevant documents represented by a particular recall value. (Recall of 0.1 may represent 10 documents in the baseline and 100 in the VLC.) If the precision-recall curve remains the same for both collections then precision at a fixed number of documents retrieved must increase.

Hypothesis 4: Swets [24] postulated separate distributions of document scores for relevant and irrelevant documents (for a given combination of query, collection and retrieval system). He assumed that the two distributions were normal and that their means differed by an amount M which could be used to characterise the performance of the combination of query, collection and retrieval system. If the distributions are the same for the VLC and for the sample, then taking either a fixed score cutoff or a fixed recall cutoff will result in the same precision - but very many more documents in the VLC than in the sample. In order to retrieve a fixed *number* of documents, the cutoff score must be set much higher in the VLC than in the sample, resulting in a greater proportion of relevant documents among those retrieved.

Hypothesis 5: The performance of retrieval systems relying on $tf.idf$ models may be harmed by anomalous df values. (Here, tf is the number of times a query term occurs in a document and idf is the reciprocal of the number of documents in the collection which contain the term. A simple method for scoring the relevance of a document involves summing $tf.idf$ values for each of the query terms, assuming that terms which occur in many documents are less useful indicators of relevance, and that the more occurrences of a query term there are in a document the greater the likelihood of its relevance. All of the systems used in the VLC track, excepting that of the University of Waterloo share these assumptions and make use of idf or a similar collection-derived frequency.) It is possible that df values obtained from a very large collection are more generally useful. From this, we might expect better retrieval effectiveness over the sample collection if dfs from the full collection (scaled down, if necessary) were used. This hypothesis does not explain the difference observed by the University of Waterloo, whose retrieval system did not use df information.

As pointed out at the beginning of this section, the above hypotheses are not necessarily mutually exclusive. A follow-up paper will analyse their relation to one another and empirically test their validity.

An explanatory model which allowed reasonably accurate prediction of the extent of growth in early precision for a given collection scale-up, could be used to control the *degree of optimization* applied when processing a query. By "degree of optimization", is meant the extent to which processing short-cuts are applied (for example, neglecting to process apparently low-value query terms) in order to increase speed, possibly at the expense of effectiveness. It may be that acceptable early precision is obtainable using quick but unsophisticated algorithms, provided that the collection is large enough.

Alternatively, the same model could be used to determine what size random sample is needed to achieve early precision results which are a specified percentage of those obtainable over the full collection. In an environment in which documents were randomly distributed across servers, size of sample would translate directly into number of servers to select.

7. Status of the Proposal for a Second-Edition VLC

The TREC-6 VLC track results clearly demonstrate that there are a number of TREC-oriented retrieval systems for which query processing over 20 gigabytes is quite feasible.

Good retrieval performance over 20 gigabytes does not demand the use of exotic and expensive hardware. The best evidence for this conclusion is provided by the Waterloo run using a US$7440 cluster of four commodity PCs, connected by a 10 Mbit/sec. network. 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.

The only apparent downside to the method used was the amount of disk space required.

Unfortunately, it is now clear that the one-order-of-magnitude increase in collection size represented by the first edition VLC does not go far enough. Almost all of the TREC-6 VLC participants indicated a desire to attempt query processing over a 100 gigabyte (10^{11}) byte collection.

Accordingly, a second-edition VLC (known as VLC2) has been constructed using data from the World Wide Web collected by the Internet Archive [17] and distributed by ACSys [15] to TREC-7 VLC track participants. It comprises 100 gigabytes of data, and two uniform subsets, one of 1 gigabyte (1%) and the other of 10 gigabytes (10%), have been defined.

Guidelines in the TREC-7 VLC track are very similar to those of TREC-6 and the track focus is on exploring the effects of two orders of magnitude scaling of data. At the time of writing, assessment of TREC-7 VLC track submissions was under way.

8. Conclusions and Future Work

The 20 gigabyte first-edition Very Large Collection appears to be a useful adjunct to the popular TREC test collection. It adds new data sets which may be interesting in their own right. It allows TREC participants to qualify their systems on a quantity of data which is a reasonable fraction of the size indexed by leading Web search services. It allows

performance data to be obtained at the 20 gigabyte level and also permits retrieval system developers to measure the scalability of their system over a one order of magnitude increase in collection size.

The recently defined VLC2 extends the test collection to a size representative of all but the very largest current applications. As such it may allow meaningful comparison of state-of-the-art commercial and research systems. Performing retrieval over 100 gigabytes of data will be a considerable challenge for TREC participants, particularly for those using a single machine. System implementors who succeed at this level will gain very useful information about the scalability of their system and a good indication of how it will perform at the terabyte level.

Remembering that the text holdings of the Library of Congress (in paper form) are estimated to reach 17 terabytes, it does not seem unrealistic to expect a number of future real applications to involve collections of this magnitude. Indeed, as discussed in the Introduction, the Alta Vista search engine indexes approximately one terabyte of text and the combined Lexis-Nexis on-line databases approach four terabytes. [8, 18]

Participants in the TREC-6 VLC track who used the same hardware for baseline and 20 gigabyte runs observed an essentially linear relationship between query processing time and collection size. Such a relationship may not be acceptable in practice as, for example, 3.6 seconds to process a query over a gigabyte would grow to one hour for a terabyte. Growth in query processing times may be reduced if hardware is scaled up but perhaps it may be possible to find optimizations or sub-collection-selection strategies which maintain effectiveness and demonstrate sub-linear scaling on a fixed hardware configuration.

The fastest query processing times and the best early precision on the TREC-6 VLC task were achieved by the shortest queries. These queries were manually generated. If similar-length, similarly well-performing queries could be generated automatically, the degree of improvement required of subsequent optimization would be significantly reduced.

Results in the TREC-6 VLC track suggest that clusters of very cheap workstations are capable of providing rapid query processing times over large collections. However, it is possible that high query loads may improve the relative cost-effectiveness of large-memory SMP systems. Both these types of architecture, and others, are currently employed by large-scale commercial service providers [16, 7, 18] but, due to lack of complete information, it is difficult to draw meaningful conclusions about their relative merits.

Further work is needed to determine whether time-sharing of multiple queries can be made advantageous in terms of throughput as well as of short-query waiting-time.

Hopefully, the existence of a static collection of realistic size and the use of efficiency as well as effectiveness measures in the context of the VLC will encourage some Web search engine developers to calibrate the performance of their systems (on both dimensions) within the standardised framework of TREC.

The immediate challenges for the VLC community are to test the hypotheses about why early precision increases as collection size grows and to seek retrieval algorithms which combine high effectiveness with high speed over collections of 100 gigabytes and larger. It seems likely that systems operating over the largest collection sizes will need to include re-ranking algorithms to avoid showing the user documents identical, or very similar, to documents already presented.

Acknowledgements

Mark Sanderson and Jon Ritchie of Glasgow University arranged for the release of data from the Glasgow Herald and for additional data from the Financial Times. Gordon Cormack and Rob Good of the University of Waterloo supplied huge quantities of archived USENET news. Ellen Voorhees and Dawn Tice (Hoffman) of NIST provided advice, support and practical assistance. Jason Haines, Tim Potter and Nick Craswell of ACSys formatted the new data and Deborah Johnson, Sonya Welykyj and Josh Gordon assessed TREC-6 submissions.

We are indebted to the above individuals and particularly to the organizations who gave permission to include their data in the VLC. VLC donor organizations 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).

Finally, thanks are due to all the participants in the TREC-6 VLC track who overcame substantial hurdles and completed the task. Without their efforts, there would have been no results to analyse or explain and no indication of state-of-the-art. Their enthusiasm to increase the collection size to 100 gigabytes testifies either to a very strong desire for understanding, . . . or to masochism!

Notes

1. Reportedly 36 million per day in the case of Alta Vista as at May, 1998
2. User-friendly software tools have been developed to support relevance assessment. Groups interested in gaining access to the tools are encouraged to contact the first author.
3. Contact the first author for details of availability.
4. In the UMass case, the processor was one of four in an SMP system.
5. IBMs used a similar approach but added all the individual times to produce a single-system time.
6. However, the consequences do not appear to have been too drastic, as VLC results show an appreciable increase in Precision@20 over the baseline, where it is understood that un-normalised fusion was not used.
7. Some search services actually comprise a number of replicated search engines (each with their own copy of index data). In this discussion, replication is ignored and queries arriving at a single engine are considered.
8. p. 173

References

1. Alexa Corporation. Web spawns 1.5 million pages daily. `http://www.alexa.com/company/inthenews/webfacts.html`, August 1998. Press release.
2. J. Allan, J. Callan, W.B. Croft, L. Ballesteros, D. Byrd, R. Swan and J. Xu. INQUERY does battle with TREC-6. In Voorhees and Harman [25], pages 169–206. NIST special publication 500-240.
3. E.W. Brown and H.A. Chong. The GURU system in TREC-6. In Voorhees and Harman [25], pages 535–540. NIST special publication 500-240.
4. C. Buckley and A. F. Lewit. Optimisation of inverted vector searches. In *Proceedings of the Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 97–110. ACM, New York, 1985.
5. C. Buckley, A. Singhal and M. Mitra. Using query zoning and correlation within SMART: TREC-5. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*, pages 105–118, Gaithersburg MD, November 1996. U.S. National Institute of Standards and Technology. NIST special publication 500-238.
6. G.V. Cormack, C.R. Palmer, S.S.L. To and C.L.A. Clarke. Passage-based refinement: Multitext experiments for TREC-6. In Voorhees and Harman [25], pages 303–320. NIST special publication 500-240.
7. Digital Equipment Corporation. About Alta Vista web page. `http://www.altavista.digital.com/av/content/about.htm`, 1997.
8. Digital Equipment Corporation. Digital's Alta Vista search index grows to record heights. `http://www.altavista.digital.com/av/content/pr052798.htm`, May 1998. Press release.
9. M. Franz and S. Roukos. TREC-6 ad hoc retrieval. In Voorhees and Harman [25], pages 511–516. NIST special publication 500-240.
10. Free Software Foundation. GNU WGET manual. `http://theory.uwinnipeg.ca/localfiles/infofiles/wget.html`, 1997.
11. D. K. Harman, editor. *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.
12. D. Harman. Overview of the fourth Text Retrieval Conference (TREC-5). In D. K. Harman, editor, *Proceedings of the Fourth Text Retrieval Conference (TREC-4)*, pages 1–24, Gaithersburg MD, November 1995. U.S. National Institute of Standards and Technology. NIST special publication 500-236.
13. D. Hawking. Scalable text retrieval for large digital libraries. In Carol Peters and Costatino Thanos, editors, *Proceedings of the First European Conference on Digital Libraries*, volume 1324 of *Lecture Notes in Computer Science*, pages 127–146, Pisa, Italy, September 1997. Springer, Berlin.
14. D. Hawking, P. Thistlewaite and N. Craswell. ANU/ACSys TREC-6 experiments. In Voorhees and Harman [25], pages 275–290. NIST special publication 500-240.
15. D. Hawking, P. Thistlewaite and N. Craswell. *TREC Very Large Collection (VLC) web page*. ACSys Cooperative Research Centre, The Australian National University, Canberra, 1997. `http://pastime.anu.edu.au/TAR/vlc.html/`.
16. Inktomi Corporation. The Inktomi technology behind HotBot. A white paper. `http://www.inktomi.com/Tech/CoupClustWhitePap.html`, 1996.
17. Internet Archive. Building a digital library for the future, August 1997. `http://www.archive.org/`.
18. Lexis-Nexis Corporation. About Lexis-Nexis web page. `http://www.lexis-nexis.com/lncc/about/datacenter.html`, September 1998.
19. National Institute of Standards and Technology. TREC home page. `http://trec.nist.gov/`, 1997.
20. V. Paxson. Flex documentation for linux. `http://www.elcafe.com/man/man1/flexdoc.1.html`, 1991.
21. M. Persin. Document filtering for fast ranking. In W. Bruce Croft and C.J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 339–348, Dublin, Ireland, July 1994. Springer, Berlin.
22. G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.
23. A. Singhal. AT&T at TREC-6. In Voorhees and Harman [25], pages 215–226. NIST special publication 500-240.
24. J.A. Swets. Information retrieval systems. *Science*, 141(3577):245–250, July 1963.
25. E. M. Voorhees and D. K. Harman, editors. *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*, Gaithersburg MD, November 1997. U.S. National Institute of Standards and Technology. NIST special publication 500-240.

26. S. Walker, S.E. Robertson, M. Boughanem, G.J.F. Jones, and K. Sparck Jones. Okapi at TREC-6: Automatic ad hoc, VLC, routing, filtering and QSDR. In Voorhees and Harman [25], pages 125–136. NIST special publication 500-240.
27. L. Wall, Tom Christiansen, and Randal L. Schwartz. *Programming Perl*. O'Reilly and Associates, Sebastopol CA, 1996.