# Coping with Growing Collections of Electronic Text

David Hawking
Co-operative Research Centre for Advanced Computational Systems
Department of Computer Science
Australian National University
dave@cs.anu.edu.au

Kerry Webb
Co-operative Research Centre for Advanced Computational Systems
CSIRO Centre for Mathematics and Information Science
Kerry.Webb@cmis.csiro.au [*]

February 19, 2007

## Abstract

Despite the trend toward distributed information sources, future digital libraries may hold as much text in electronic form as current libraries do in print. Accessing such collections by content rather than by metadata will require search-engine technology to accommodate at least a hundred-fold growth in data size. Recent developments within the ACSys Cooperative Research Centre are described, including an effective and cost-effective retrieval system (PADRE) designed to scale to multi-terabyte levels, a very large test collection for use in retrieval evaluation, techniques for selecting information servers and combining their results and ideas for combining content searches with access by metadata.

[*] 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 Centre's Program.

# 1 Introduction

As illustrated in figure 1, the text holdings in conventional libraries still dwarf the quantity of text available electronically. Using a "book" (where 1 book = 1 megabyte) as the unit of size for text collection collections, the U.S. Library of Congress holds more than 17 megabooks, about 34 times as much text as currently available on the World Wide Web (estimated at 500 kilobooks). Searchable electronic text collections are even smaller. Currently, the largest Web search engine indexes only of the order of 50 kilobooks and, even in large institutions, retrieval systems typically operate over only a few kilobooks worth of data.

Resolution of a number of political, social, economic, technological and legal issues will determine the form and role of future libraries, if and when paper collections are supplanted or significantly supplemented by electronic ones. Similar issues may also profoundly affect the future of bookshops, academic journals and of publishers.

On the technological front, there is uncertainty at present:

1. about the extent to which distributed information retrieval will replace the need for centralised collections; and

2. about the extent to which text documents will be located via content using text retrieval-style operations rather than via metadata (i.e. cataloguing information) using database-style operations eg.

```
select *
from book
where author = 'Arthur C. Clarke'
and year < 1990 and year > 1980
```

There is no doubt that future readers will access distributed sources of information to a greater extent than previously possible in traditional libraries. Furthermore, considerable effort is now being invested in developing metadata standards and in creating metadata descriptions of electronic materials. However, it is the view of the present authors that a considerable proportion of accesses to at least the non-fiction holdings of future digital libraries will be mediated by content and that the size of the electronic text holdings of at least some of these libraries will grow to the multi-megabook range.

Consequently, there will be a need for text retrieval methods which:
- are fast;
- scale from the current kilobook level to the megabook level;

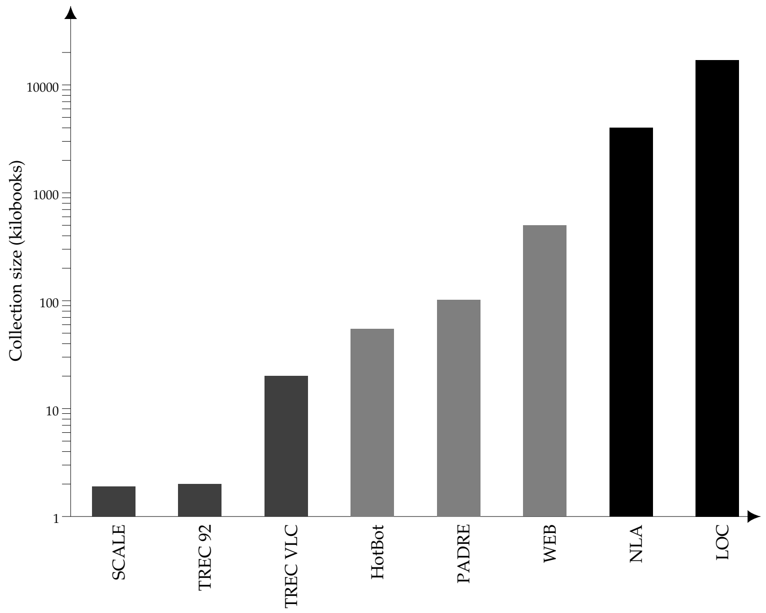

Figure 1: Comparison of various text collection sizes. (Note logarithmic scale.) Bars marked TREC and TREC VLC represent the sizes of two test collections available for evaluation of search engines. SCALE represents the amount of text in the SCALE legal databases operated by the federal Attorney-General's department. HOTBOT represents the largest amount of text currently indexed by a Web search engine (estimated to be 10% or less of the total amount of text on the WEB.) PADRE shows the largest amount of text so far used in experiments with the current version of the ACSys/ANU PADRE search engine. NLA and LOC show estimates of the amount of text data in the paper collections of the National Library of Australia and the U.S. Library of Congress respectively.

- operate on cost-effective, scalable hardware;

- do a good job of retrieving what the user wants; and

- can be combined effectively with database-style operations over metadata.

This paper describes a text retrieval system developed in the ACSys Cooperative Research Centre which is competitive with leading international systems on both speed and effectiveness when run over kilobook collections on desktop hardware. The same system has been shown to to process reasonably complex search queries over a replicated collection whose size exceeded 100 kilobooks.

Future digital libraries will almost inevitably contain large-scale holdings of video, images, sound and three-dimensional electronic artefacts as well as text. Methods discussed here may be relevant to the retrieval of such items using textual captions, transcriptions or metadata. However, retrieval using non-textual properties is beyond the scope of the present work.

## 2 Evaluating Text Retrieval Systems

Performance (deliberately ignoring user-interface, compatibility and "feature" aspects) of text retrieval systems is generally measured along three axes:

1. Indexing speed. How long does it take to (automatically) build the indexes and other data structures necessary to process queries? How long does it take to respond to deletions, insertions and alterations to the document collection? Depending upon the pattern of changes to the collection, these questions may be vitally important or largely irrelevant.

2. Query-processing speed. How long does it take to process a user query? How many queries can be processed per hour? Both measures may be important.

3. Query-processing effectiveness. Recall: of all the documents in the collection which are actually relevant to the user's information need, what proportion can be retrieved by the combination of query and retrieval system? Precision: of the documents retrieved by the system, what proportion are actually relevant? Average Precision is a single-number measure which takes into account both precision and recall.

Measurement of indexing speed requires only the availability of a standardised collection and [perhaps] a standardised hardware platform. Measurement of query processing speed has the same requirements and in addition requires standardised queries. However, neither of the speed measures is of much interest in the absence of effectiveness data. After all, it's easy to retrieve items quickly if it doesn't matter what you retrieve! Effectiveness data relies on the availability of standard retrieval problems and corresponding "correct answers".

In the early days of information retrieval research, effectiveness experiments were conducted with very small (one or two books) homogeneous collections of short documents (e.g. Cranfield and CACM collections). A standard set of queries was devised for the collection and experts read ALL of the documents in the collection for relevance to each of the queries.

Useful results were obtained using these tools but the collections were soon seen to be too small relative to the data collections typical of real applications. Accordingly, in 1992 the U.S. National Institute of Standards and Technology (NIST) with support from the Defense Advanced Research Projects Agency (DARPA) established a series of Text Retrieval conferences (TREC, [**?**]) based around a 2 kilobook collection (over half a million documents) from a variety of sources such as patents, government documents and newspaper and technical magazine articles. Participation in specified retrieval tasks over the TREC collection is a requirement of admission to the conference.

With a collection of the TREC size, it is not practical to read the entire collection for relevance. At a rate of two books per day, it would take about eight working years per research topic! Accordingly, the documents retrieved by a large number of [hopefully] diverse automatic retrieval systems are assumed to be a superset of the relevant documents. These documents are judged by humans and unjudged documents are deemed to be irrelevant.

The TREC conference is an annual event and draws a large and growing number of participants from among commercial text retrieval companies (such as Verity, Fulcrum, Excalibur, SABIR and OpenText), search engine companies (Excite), document service providers (such as Lexis-Nexis and Westlaw), government departments (such as the CIA and GCHQ), hardware vendors (such as IBM, Apple, NEC, Xerox and AT&T) and universities (such as Cornell, City, Dortmund, Berkeley, ANU and RMIT/Melbourne).

Each TREC research topic is an English-language description of someone's information need expressed in a form suitable to be given to a human assistant. An example is shown in figure 2. Different categories of participation allow manual or one of several types of automatic conversion of the research topic into queries for the participating system.

```
<num> Number: 301
<title> International Organized Crime

<desc> Description:
Identify organizations that participate in international criminal
activity, the activity, and, if possible, collaborating organizations
and the countries involved.

<narr> Narrative:
A relevant document must as a minimum identify the organization and the
type of illegal activity (e.g., Columbian cartel exporting cocaine).
Vague references to international drug trade without identification of
the organization(s) involved would not be relevant.
```

Figure 2: An example TREC research topic, taken from TREC-6.

After five TRECs there are now relevance judgments for 300 different topics. The TREC materials thus comprise a unique and highly valuable resource for improving retrieval systems. Large sets of queries can be repeatedly run and evaluated for effectiveness using the standard TREC evaluation package while varying query generation, indexing or query processing parameters.

Within guidelines of fairness and proper scientific method, the TREC framework may be used to compare the effectiveness of retrieval systems, but it is important to note that TREC submissions vary enormously in the amount of human and machine resources committed and in the purpose of the submission. Some runs are submitted with the aim of testing a particular hypothesis rather than "winning".

## 2.1 Evaluating Performance on Larger Collections

In 1997, the ACSys Cooperative Research Centre, with the cooperation of NIST have created and distributed a new 20 kilobook research collection and organised a track within the TREC conference devoted to studying the scalability of retrieval systems as the data size is increased from 2 kilobooks to 20.

The new collection (called VLC for Very Large Collection) includes several kilobooks of data from Australian institutions such as the federal Attorney-General's Department, the Parliament of Australia, the Federal Department of Industrial Relations, the Australian Broadcasting Commission, the National Library of Australia, the Commonwealth Industrial and Scientific Research Organisation, the Department of Defence, the Australian Computer Society and ten Australian universities as well as data from the U.S., England, Scotland and Canada.

Participants in the VLC track at TREC must generate queries for the 50 TREC-6 topics and submit their 20 top-ranked documents for each query over both the 20 kilobook collection and a uniform 10% sample. Indexing speed, query processing speed and retrieval precision are computed for both the full collection and the sample. Speed ratios are used as a measure of scalability.

At the time of writing, relevance assessors employed by ACSys are reading all the documents submitted and assessing their relevance. Once the assessments are complete, precision figures can be calculated. Unfortunately, because of the scale of the collection, it is not feasible to measure recall. Indeed, the VLC task is of most use as an adjunct to the main TREC task, system performance being measured on the main task and scalability to larger collections being confirmed with the VLC. The VLC precision measure is mainly used to prevent participants achieving scalability at the expense of effectiveness.

# 3 Scalable Text Retrieval - the PADRE System

The current version of the PADRE retrieval system (PADRE97) is described in ?]. The same paper compares PADRE with the leading TREC AdHoc (2 kilobook collection) participants on the three performance measures listed above and concludes that PADRE is capable of performing at or near state of the art levels on all three. It should be noted that the PADRE effectiveness results are unofficial and were obtained recently rather than at the time of the particular TRECs. However, the automatic algorithms were tuned on one set of TREC data and tested on others. Improvements to relevance feedback algorithms and the introduction of query optimisations may improve results even further.

?] reports the results of experiments designed to test scalability of query-processing time with increasing data size, using both single workstations and clusters of such workstations. Further results are reported in ?]. Hardware used included Fujitsu, Sun, DEC and SGI systems.

At best, query processing speed on a single workstation was found to

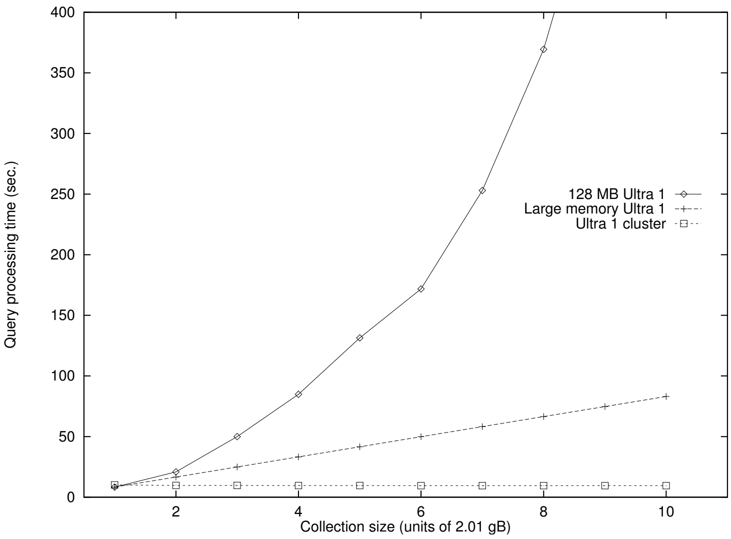

Figure 3: Elapsed query processing times for processing collection sizes measured in 2.01 gigabyte units on three different systems: a 128 MB Sun Ultra 1 (observed times); the same Ultra 1 with memory hypothetically increased in proportion to the data size (projected times); and a cluster of Ultra 1s with one search engine for each unit of collection size (scaled up from the data obtained while processing smaller datasets using a cluster of slower SPARC systems.)

exhibit a linear deterioration with increasing data size. At worst, increasing virtual memory paging resulted in dramatic degradation. By contrast, it was found that if the number of workstations in the cluster was increased in proportion to the size of the data, query processing time remained constant, provided that the collection was evenly distributed over disks local to the workstations and provided that the search manager was not running on the same workstation as a search engine. These results were obtained for small collections on up to 15 workstations connected by a low-latency network and for large collections on up to ten workstations connected by a high-latency network (Ethernet).

Queries were successfully processed over 102 kilobooks of data using ten workstations, each with only 64 MB of RAM.

Scalability results obtained are illustrated in Figure 3 which has been reprinted from [?].

# 4 Distributed Information Retrieval

Despite the likelihood of very much larger text collections, users will increasingly be involved in retrieving information from geographically distributed sources. The factors leading to the creation, growth and heavy use of the world-wide web are very unlikely to be reversed. Accordingly, research within ACSys is also being conducted on distributed information retrieval.

The centralised web index model exemplified by Alta Vista and HotBot suffers from three deficiencies: poor coverage (probably less than ten percent of text on the web is indexed), slow response to changes in web data and lack of quality control on documents indexed.

These deficiencies could be alleviated if an automatic search manager maintained a list of authoritative, searchable sites (such as digital libraries) and forwarded user queries to a suitably selected subset, merging results for presentation to the user. To be successful however, a number of problems must be solved:

1. Selection of a subset of servers in a way which achieves the desired results while avoiding excessive network and searching costs.

2. Translation of the user query into the diverse query languages of the selected servers.

3. Combining the results into an optimal ranking for presentation to the user. This is much more difficult than it seems.

?] have proposed a dynamic method for server selection based on lightweight probe queries and shown that it is competitive with other methods reported in the literature. In other papers [?; ?] distance-based (rather than the ubiquitous vector-space and probabilistic methods which rely only on frequencies) relevance scoring is shown to solve the problem of merging rankings and to be capable of achieving high effectiveness results given suitable queries.

# 5 Combining Content Searching with Metadata Matching

Combining text retrieval operations with database-style operations is a natural thing to do. For example: “Find me documents about repression of minority groups in the Middle-East which have been published since 1994”, or “I'm looking for documents published by the American Society for Information Science which deal with intellectual property issues relating to digital libraries.”

Two potential approaches to supporting these combined operations will now be described.

## 5.1 Grafting Text Retrieval Operations onto a DBMS

Many database vendors allow variable length text to be stored and indexed (using text indexing methods) within particular columns of a relational database. SQL commands can then be used to select records which match the necessary attribute conditions and rank the resulting subset according to the content of the text field or fields.

Textbook SQL examples such as the one given in the introduction to this paper give the impression that testing attribute conditions is a trivial matter but they assume that attribute values are consistently, accurately and unambiguously recorded and are universally present. In practice, missing attribute values, inconsistent and ambiguous recording of values and differing representation formats are a pervasive fact of life. They inevitably give rise to problems, either in creating a “clean” database or in dealing with an “unclean” one. For example, if no publication date is specified for a document, should it be taken to satisfy the “after 1994” condition, or not? Should “Arthur Clarke” or “Clarke, A.C.” be considered to match “Arthur C. Clarke”? What about “Arthur C. Clark”? Can “9/10/97” and “le neuf octobre, mille huit cent quatre-vingt dix-sept” be regarded as equivalent? Should either of them be considered to satisfy the “after 1994” condition?

## 5.2 Grafting Attribute Conditions onto a Text Retrieval System

It is also possible to start with a content-based search engine and to add database-like functionality. The complete text of each document must be scanned during data structure building. During this scan, it would be quite feasible to recognise XML (for example) tags giving the necessary attribute information and to record attribute values in extra fields added to the document table. The query language may also require extension to make use of the attribute information.

The resulting enhanced text retrieval system might mimic the operation of the textually enhanced DBMS by performing ranking only on the documents found to satisfy the attribute conditions. Additional data structures could be built to speed up access to document attributes.

Alternatively, all documents might first be ranked according to content but then be presented to the user in groups defined by the extent to which they satisfy the attribute conditions, first those which definitely satisfy all conditions, then those with one missing value, etc. Work on multi-valued logics is relevant here.

As a final alternative, attribute matching could be incorporated into the weightings used to rank documents rather than being applied as a strict filter.

# 6 Conclusions

Digital libraries of the future are likely to contain volumes of electronic text much greater than the largest of those indexed by current search engines. In such a context, retrieval effectiveness, and the TREC framework for evaluating it, will become even more important than at present. Recent ACSys research has resulted in both a retrieval system (PADRE) designed to scale to the necessary level and a test collection and methodology for evaluating scalability. PADRE has been tested on datasizes in excess of 100 gigabytes, using a cluster of relatively low cost workstations.

Future information-seekers should be given the capability to retrieve documents through a combination of their metadata attributes and their textual content. A number of approaches to this problem have been discussed but require further evaluation. Recognizing that future digital libraries will be only a part of a highly distributed information smorgasbord, ACSys research has also adressed perceived shortcomings of the present centralised web indexing services and led to promising methods for server selection and result merging.

# Acknowledgements