A New Zealand digital library for computer science research

Ian H. Witten, Sally Jo Cunningham, Mahendra Vallabh
Department of Computer Science
University of Waikato
Hamilton, New Zealand
email: ihw@waikato.ac.nz; sallyjo@waikato.ac.nz; mike@waikato.ac.nz

Timothy C. Bell
Department of Computer Science
University of Canterbury
Christchurch, New Zealand
email: tim@cosc.canterbury.ac.nz

Abstract: A large amount of computing literature has become available over the Internet, as university departments and research institutions have made their technical reports, preprints, and theses available electronically. Access to these items has been limited, however, by the difficulties involved in locating documents of interest. We describe a proposal for a New Zealand-based index of computer science technical reports, where the reports themselves are located in repositories that are distributed world-wide. Our scheme is unique in that it is based on indexing the full text of the technical reports, rather than on document surrogates. The index is constructed so as to minimize network traffic and local storage costs (of particular importance for geographically isolated countries like New Zealand, which incur high Internet costs). We also will provide support for bibliometric/scientometric studies of the computing literature and our users.

1. Introduction

The migration of information from paper to electronic media promises to change the whole nature of research, and in particular the nature of indexing and how information is located. Small, geographically isolated countries like New Zealand, which find the struggle to keep up with the explosion of printed information exceptionally difficult, stand to benefit greatly from networked electronic libraries. Indeed, in some fields (notably physics) this process has begun already, as researchers in less developed report access to ongoing research through the Internet that their local libraries could not afford to acquire through conventional journals [Ginsparg, 1994a-b]

This paper describes a pilot project that is designed to explore the potential of such digital libraries in the context of a national research community that is relatively small and focused. The project will provide a full-text index to computer science technical reports accessible via Internet, and make it available to academic researchers in New Zealand computer science departments. Two factors that are heightened by geographical isolation are network transmission costs and response time variability. To reduce trans-Pacific network traffic and local storage requirements, the full text of the technical reports will not be transferred but downloaded on demand. Particular attention is being paid to scalability, and it is intended that growth be socially controlled in a manner described below.

By far the greatest problem in setting up a digital library is in obtaining and digitizing the raw material. Computer science is an appropriate choice for a digital library

because a huge amount of high-quality information already exists in digital form and is freely accessible on the Internet in the form of technical reports. For example, the University of Indiana maintains a *Unified Computer Science Technical Report Index* which, as of October 1994, contained 10,500 items found at 180 Internet sites (an estimated 10 gigabytes of data). The list of sites is growing quickly and is potentially much larger.

Our design goals for the system include:

- *Low maintenance*. The index construction process will be automated as much as possible, and documents will be added to the index with little or no intervention from a system administrator.

- *Logically central index, physically distributed documents*. The New Zealand site will hold only an index and search engine; the documents themselves remain in their original repositories.

- *Full text indexing*. The system will index the entire contents of the documents, rather than being restricted to file information or title/author/abstract summaries.

- *tTansparency for providers*. The system will not require any effort on the part of participating technical report repositories, and indeed these providers will in general not even be aware of their inclusion in our index. No special software, archive organizations, or file formats will be required of the providers.

This paper is organized as follows: Section 2 compares our proposal with existing technical report indexing schemes; Section 3 describes the proposed system architecture in detail; Section 4 briefly discusses the potential for gathering scientometric/bibliometric information on New Zealand researchers and the computer science literature; and Section 5 presents our conclusions.

2. Other technical report indexing systems

A number of technical report searching and indexing systems currently exist, including: the physics E-PRINT ARCHIVE, which has supplanted journals or pre-print mailings as the primary information dissemination point for several areas of physics [Ginsparg, 1994a-b]; the above-mentioned Unified Computer Science Technical Report Index (UCSTRI) in Indiana; the NTRS system, which provides an index for NASA technical reports [Nelson, 1994]; the WATERS distributed database of computer science technical reports [Maly, 1994]; the DIENST repository, indexer, and search engine that currently provides an interface to a handful of computer science technical report repositories [Davis, 1994a-c]; Carnegie Mellon's MERCURY search engine (again, currently applied to computing literature); and the HARVEST tools for information gathering, index construction, and searching [Bowman, 1994a-b]. A list containing these and other subject- or site-specific technical report servers is maintained at [NASA].

The provision of a full-text index of the entire contents of each document is unique to our proposed system. Other schemes index on user-supplied document descriptions, abstracts, or similar document surrogates. UCSTRI, for example, provides a searchable text index based on text obtained by parsing the index file that is present by convention in most ftp directories of technical reports. This text does not necessarily characterize the report very closely, and in any case is only a small subset of the full text in the document. Moreover, the parsing procedure is sensitive to the format of the index file, and cannot be guaranteed to succeed. HARVEST's ESSENCE sub-system [Hardy, 1993; Hardy, 1994] extracts "content summaries" whose composition may vary widely. ESSENCE relies on filetype-specific procedures to extract relevant information from the document itself; for example, LaTeX documents can be parsed for author and title information. ESSENCE 's success in extracting an appropriate document surrogate

depends on its ability to cope with the file type of the document and the semantic cues provided by that type.

The remaining systems under consideration – DIENST, NTRS, WATERS, and the physics e-print archives – require the submitter of the technical report to provide cataloging information, while WATERS requires a designated site librarian to maintain a local catalog.

UCSTRI, HARVEST, and our proposed system primarily provide keyword searches, as their indices do not contain formal bibliographic catalogs. The DIENST, NTRS, WATERS, and physics E-PRINT ARCHIVES can support more detailed information about each report (such as author, title, and CR category field searching), but this more sophisticated search functionality comes at the expense of requiring participating repositories to use specific software. As a consequence, these latter systems provide access to only a handful of sites, whereas UCSTRI, HARVEST, and our system can access a broad range of providers.

3. The project

The system will provide a full-text index to this raw material, stored locally on an Internet node in New Zealand. In addition, an automatically-extracted abridgment of each document, containing material such as the first page, will be held locally, along with a pointer to the remote ftp site and directory where the original resides. The index will allow full-text access by Boolean search or ranked search to all technical reports. Queries will be answered very quickly, and will result in the abridged document being presented. This can be examined and, if desired, the entire document can be downloaded from the remote site (provided it has not been moved or deleted).

The scheme rests on the feasibility of full-text indexing of large corpora of text. The public-domain system *mg* can store a full-text index to a large collection of text in only 5% of the size of the original text [Witten, 1994]. Further, it provides a search engine that can process queries efficiently. Experiments with the 750,000 document TREC collection give response times of three to five seconds to produce ranked output for queries of forty to fifty terms.

Computer science departments generally make their technical reports available in PostScript form, although our system could support other formats such as DVI and RTF should this prove necessary. Software is available for extracting plain text from such files, and this text can be indexed to allow searching. Some sites (for example, Cornell University) provide backward compatibility to their non machine-readable technical report archives by storing old reports as TIFF files of page images, together with an ASCII version of the text (obtained by OCR) [Davis, 1994]. These can be incorporated naturally into our scheme, with the ASCII text being indexed and the page images retrieved when the report is requested.

With the cooperation of a site geographically close to the text archive being indexed, there is no need for the entire remote archive to be transmitted. Only the vocabulary list and word count, along with the abridged version of the document, is needed to update the local information base. A typical technical report occupies 1 megabyte in PostScript form, or 250 kilobytes compressed, while the compressed indexing information (a list of words and their frequencies) occupies only 10 kilobytes. The cooperating site will download all the documents, strip out the text, and send the appropriate information to the New Zealand host. This activity can be carried out when the machines and the network are lightly loaded. Given the currently high Internet costs to New Zealand, this scheme will be significantly less expensive than downloading directly to New Zealand and performing the indexing there. If sufficiently many sites cooperated, new indexing information could be exchanged systematically in much the same way that the Internet news is propagated.

The search interface

As noted above, the *mg* search engine supports Boolean keyword searches on the full text of documents. Since the system will not have access to any formal cataloging information about the technical reports, it cannot support descriptive field searching (such as author, title, and publication date). Instead, we will provide the following types of search delimiters:

- *Timestamp*: A coarse type of publication date search is supported by specifying a desired range of dates in which the technical report was entered into its repository. Given that a number of repositories are digitizing their older paper reports, this type of search is likely to produce uneven results (since the timestamp can only record the data that the report was inserted into the database, not the date that it was originally produced). However, we expect timestamp search to become more accurate as the repositories "catch up" on their retrospective conversion.

- *Initial page*: In the vast majority of reports, the first page contains important bibliographic information such as the title, author, the institution with which the author is affiliated, etc. By limiting a search to this first page, the user can approximate a search based on this type of information (for example, an initial page search for documents authored by "Knuth" will not retrieve documents which only cite previous work by him).

This approach avoids requiring system administrators or report providers to provide formal cataloging for technical reports, in accordance with our goal to eliminate the need for active participation from repositories. An intelligent document parser such as HARVEST 's ESSENCE summarization system [Hardy, 1993; Hardy, 1994] could potentially provide more precise bibliographic information, but at the expense of requiring a significantly more complex and filetype-dependent indexing system.

Ranked query output

Exhaustive document indexing has been shown to improve query recall (the proportion of relevant documents retrieved) at the expense of query precision (the proportion of retrieved documents that are actually relevant). By ranking search results so that documents more likely to be relevant are presented to the user first, a system gains increased retrieval precision and greater user satisfaction [Salton, 1983].

Mg uses standard techniques for producing ranked query output. A list of index terms and term frequency statistics from the documents in the collection are used to assign weights to terms. Using the vector document representation, similarity between a query and document can be measured as the cosine of the angle between their two vectors. The user then need only provide a list of words relevant to the topic of interest, and the system automatically ranks documents according to their "closeness" to the query terms.

Size and scalability

The ASCII text (not pictures) of the 10,500 technical reports currently indexed by Indiana is estimated to occupy under 1 gigabyte and could be stored in its entirety in around 350 megabytes compressed, together with a 75 megabyte compressed index. Our scheme will store just the index, and the volume of technical reports would have to increase by a factor of 100 before the index exhausted a 10 gigabyte disk, which, assuming a tripling every year (our best estimate of the current rate of growth), gives us four years. New sites can be added by various mechanisms, beginning with initial lists such as those compiled by [Blythe] and [Harris], manually (or automatically) scanning the newsgroups that announce new technical report lists, and encouraging users to

email suggested new sites to a central coordinator. This is how the collection will be managed in the first instance.

However, an all-inclusive information collection policy is basically unscalable and will become infeasible if the Internet continues to grow exponentially. One alternative is a socially-facilitated, access-dependent scheme for pruning the information base. This works by monitoring every participating user's (i.e. New Zealand computer science researcher's) access to technical reports, and in particular noting the sites that see the least use. These sites are prime candidates for removal when the size of the collection becomes unmanageable. The idea is that only potentially 'interesting' sites are included, where 'interesting' is defined as 'has been accessed by a colleague.' This means that the rate of growth of the collection, and hence the resources it consumes, is governed by the size, diversity, and level of activity of the user population rather than by the rate of growth of the bibliographic universe.

An interesting extension would be to measure the "potential for interest" for each new site to be added to the collection. Here, we would match the index terms for the proposed new site to the terms found in those sites receiving the heaviest use. If the new site does not sufficiently overlap the topics covered in the well-used site (and, by extension, does not cover areas that New Zealand computer scientists currently work in), then the site would not be included.

Maintenance

Stability is not, of course, one of the characteristics of the Internet. UCSTRI, another technical report server that stores only the document index and pointers to the documents themselves, reports "frequent maintenance problems" caused by changes in the technical report repositories that it indexes – for example, files being renamed or removed from the collection [VanHeyningen, 1994]. We are currently monitoring a large number of repositories to measure the stability of their collections. The results of this experiment will give us an indication of how problematic these changes to the underlying repositories will be.

On a more technical level, the current implementation of the indexing platform (*mg*) requires that the entire index be re-built each time a new document is added to the collection. A project at Canterbury University (Christchurch, New Zealand) has developed an extension (*mgmerge*) which allows several existing indices to be merged together, so that indexes can be accumulated incrementally [Hudson, 1995].

4. Opportunities for bibliometric/scientometric research

The new digital library will permit research on scientific information collection and use at a much finer grain than is possible with current paper libraries. Current bibliometric or scientometric research of this type must measure information use indirectly – for example, through examination of the list of references appended to published articles. By monitoring the use of our technical reports index, we will be able to measure usage directly, and to obtain a more detailed picture of how New Zealand computer scientists pursue their research. The size of the New Zealand academic computing research community is particularly appropriate for an intensive study, as it is limited to seven universities.

Surprisingly few previous reports on index or repository projects provide even a cursory analysis of the usage data they collect on their systems. As a sample of the types of analysis possible, Paul Ginsparg notes a seven day periodicity in the number of search requests made to the physics e-print archives. From this he adduces that many physicists do not yet have weekend access to the Internet (an alternative, slightly

more cynical hypothesis is that even high energy theoretical physicists take the weekend off) [Ginsparg, 1994].

In addition to monitoring user access times, we will also record the specific documents retrieved by users. Analysis of the index terms for these documents will allow us to create user and site profiles that characterize the types of research carried on in these departments.

Finally, there has to date been no attempt to seriously examine the characteristics of the computing literature as represented by the *contents* of the technical report repositories. Studies that could be supported by the information contained in our index include:

- examining the "physical" characteristics of computer science technical reports (for example, the range of the size of these documents as measured by their word count)

- determining the obsolescence rate of computing literature by analyzing the range of dates in the references of technical reports

- tracking shifts in the focus of individual computer science departments through analysis of changes in the terms used to index their technical reports

- detecting cycles or regularities in the rate of production of computing research, as measured by the timestamp of documents added to the repositories (for example, is more research produced over the summer, when the teaching load is lighter? or is research steadily produced throughout the year?)

5. Conclusion

This project is intended to provide, in the short term, a facility for accessing the "gray literature" contained in technical reports in the field of computer science, a field which – because the time value of information is high – relies more than most on pre-publication in the form of reports. Because of the extremely rapid rate of change in the Internet and its organization, the scheme is designed to perform a useful job in the short term, as well as lay the foundation for full-text access to substantial document collections in the future. It represents an innovative investigation into how digital libraries might benefit small, geographically isolated communities, and counter the diseconomies of scale from which they suffer in a world of exponentially growing information. The system can provide excellent response because the local indexes are small enough for each library to have a copy. This encourages browsing, and protects users from variable network loads and remote machine downtime. The project also provides a source of detailed information on information retrieval and usage by a small research community, and a platform for research on the characteristics of the computing literature as a whole.

Above all, it shows one way of dealing with the new realities of publishing, where information is provided in a widely distributed manner and it is up to the information consumer to locate what is needed.

References

Blythe, J: "On-line CS Tech reports". <URL:http://www.cs.cmu.edu:8001/afs/cs.cmu.edu/user/jblythe/Mosaic/cs-reports.html>

Bowman, C., Danzig, P., Hardy, D., Manber, U., & Schwartz, M., 1994a: Harvest: A scalable, customizable discovery and access system, *Technical Report CU-CS-732-94* , Department of Computer Science, University of Colorado, Boulder, Colorado.

<URL:ftp://ftp.cs.colorado.edu/pub/cs/techreports /schwartz/Harvest.ps.Z>

Bowman, C.M., Danzig, P.B., Manber, U., and Schwartz, M.F., 1994b: Scalable Internet resource discovery: Research problems and approaches, *Communications of the ACM 37(8)*, pp. 98-107.

Davis, J. & Lagoze, C., 1994a: Dienst, a protocol for a distributed digital document library, Internet Draft (work in progress). <URL:http://cs-tr.cs.cornell.edu/Info/dienst_protocol.html>

Davis, J. and Lagoze, C., 1994b: "Drop-in" publishing with the World Wide Web, *Proceedings of the Second International WWW Conference*, Chicago. <URL:http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Pub/davis/davis-lagoze.html>

Davis, J. & Lagoze, C., 1994c: A protocol and server for a distributed digital technical report library, *Technical Report 94-1418* , Computer Science Department, Cornell University. <URL:http://cs-tr.cs.cornell.edu/TR/CORNELLCS:TR94-1418>

Ginsparg, P., 1994a: After dinner remarks: 14 Oct '94 APS meeting at LANL. <URL: http://xxx.lanl.gov/blurb>

Ginsparg, P., 1994b: First steps towards electronic research communication, *Computers in Physics 8(4)*, p. 390-401.

Hardy, D., and Schwartz, M.F., 1993: Essence: A resource discovery system based on semantic file indexing, *Proceedings of the USENIX Winter Conference* , p. 361-374.

Hardy, D., Schwartz, M., 1994: Customized Information Extraction as a Basis for Resource Discovery, *Technical Report CU-CS-707-94* , Department of Computer Science, University of Colorado, Boulder, Colorado. To appear in an upcoming issue of *ACM Transactions on Computer Systems*. <URL:ftp://ftp.cs.colorado.edu/pub/techreports /schwartz/Essence.Jour.ps.Z>

Harris, Rik: "Computer Science Technical Reports Archive Sites." <URL:http://www.rdt.monash.edu.au/tr/siteslist.html>.

Hudson, S., 1995: Dynamic Inverted Files for Full-text Retrieval, to appear in *Proceedings of the Second New Zealand Research Students Conference* , Waikato University, April 1995.

Maly, K., Fox, E.A., French, J.C., and Selman, A.L., 1994: Wide area technical report server, *Technical Report* , Dept. of Computer Science, Old Dominion University. <URL: http://www.cs.odu.edu/WATERS/WATERS-paper.ps>

NASA: "Technical Reports, Preprints and Abstracts." (list of sites supporting digital libraries or indexing services) <URL: http://www.larc.nasa.gov/org/library/abs-tr.html

Nelson, M.L., Gottlich, G.L., and Bianco, D.J., 1994: World Wide Web implementation of the Langley Technical Report Server, *NASA Technical Memorandum 109162*, Langley Research Center, Hampton, Virginia. <URL: ftp://techreports.larc.nasa.gov/pub/techreports/larc/94/tm109162.ps.Z>

Odlyzko, A., 1995: Tragic loss or good riddance? The impending demise of traditional scholarly journals, *International Journal for Human-Computer Studies* , to

appear. Condensed version to appear in *Notices Amer. Math. Soc.*, Jan. 1995. <URL:ftp://netlib.att.com/netlib/att/math/odlyzko/tragic.loss.Z>

Salton, G., and McGill, M.J., 1983: *Introduction to modern information retrieval*, McGraw-Hill Book Company.

VanHeyningen, M., 1994: The Unified Computer Science Technical Report Index: Lessons in indexing diverse resources, *Proceedings of the Second International WWW Conference*, Chicago. <URL: http://www.cs.indiana.edu/ucstri/paper/paper.html#ref-odlyzko>

Witten, I., Moffat, A., and Bell, T., 1994: *Managing Gigabytes: Compressing and indexing documents and images*, van Nostrand.