Working Paper Series
ISSN 1170-487X

Information retrieval programs on the Internet: tools for teaching IR

by: Sally Jo Cunningham

Working Paper 95/31

November 1995

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

Information retrieval programs on the Internet: tools for teaching IR

Sally Jo Cunningham
Dept. of Computer Science
University of Waikato
Hamilton, New Zealand
email: sallyjo@waikato.ac.nz

Abstract: The theory of information retrieval has generally been taught in theory: it has been difficult to provide students with hands-on experience with retrieval engines incorporating many IR topics such as relevance ranking, fuzzy queries, etc. Recently, however, a number of retrieval programs have become freely available for interactive use over the Internet. These programs can be useful in the classroom, by permitting students to examine a variety of implementations of IR algorithms over different document collections. Moreover, many of the document collections are in themselves valuable subject resources, and are well worth exploring from the point of view of developing familiarity with them as reference materials.

Introduction

Teaching information retrieval, as opposed to training students to search specific retrieval systems, can be difficult: the theoretic results of decades of IR researchers have been slow to appear in available IR search engines, making it difficult to give students practical experience with systems including advanced retrieval features such as relevance ranking, automatic query expansion, fuzzy and non-Boolean queries, etc. While a number of good texts exist that describe these techniques and research results in detail, hands-on experimentation has been harder to provide in the classroom. The most prominent exception to this problem has been the continued availablilty of the venerable SMART system, developed by Gerard Salton and Chris Buckley at Cornell University [1]. However, this system requires a relatively large investement of effort to be successfully integrated into an IR course, as the software must be installed and maintained locally. And, of course, SMART is only one system, and does not embody all IR algorithms of interest.

In the past few months, however, a number of retreival engines have become available on the Internet, accessible through the World Wide Web. These systems can be invaluable in adding interest to an otherwise dryly theoretic subject, by giving students experience with a rich set of algorithms implemented over a variety of document collections.

A secondary, serendipitous benefit of using these IR systems is that many are based on document collections that are becoming key reference resources; familiarity with these collections is in itself an important goal for an information specialist. Some of this software is becoming part of the standard set of tools for information searches: general WWW search engines, subject collections of pre-prints and working papers, and Internet-accessible bibliographic databases.

This paper describes a selection of Internet-accessible information retrieval software, and discusses their potential uses in an IR course. We concentrate on software that can be used interactively over the WWW, as these resources are immediately useable and do not require a local investment of time or effort to support their use. However, the Appendix contains pointers to resource lists of similar software that can be downloaded for local use. Many of the programs described in this paper have only become available for use over the WWW within the past few months. While an

exhaustive inventory of IR software is not possible, given the dynamic nature of the Internet, this listing illustrates the potential that these resources provide for enriching our classes. The paper concludes with a brief discussion of some of the advantages and problems in using these resources in teaching.

Examples of available IR systems

This section will attempt to convey a flavor of the rich variety of IR systems currently available over the Internet, and to briefly discuss their potential relevance to a database/information retrieval course. URLs for these programs are presented in the Appendix.

General WWW search engines

A number of programs are available that search the World Wide Web, including Lycos, Open Text, InfoSeek, and WebCrawler. The WWW is a structure of interlinked "pages" that can contain text, pictures, and pointers to other pages or files. The pages are formatted in HTML, a hypertext markup language, and these search engines extract and index text from the Web pages. Lycos and WebCrawler restrict the indexed terms to a subset of each WWW page's full text: the title, the first few lines of a page, and certain high-information-content keywords extracted from the remainder of the document. InfoSeek indexes on the full text (up to 50K) of each page. These three systems support unfielded keyword searches, while Open Text allows users to specify that searches be limited to summaries, titles, first headings, links, and body text.

These WWW search engines are of interest to IR students on a number of counts. Their mechanisms for efficiently locating and indexing new Web pages present a good case study of effective data gathering and storage algorithms. They implement a range of query term and index term matching techniques, including weighted queries, relevance ranking, and term stemming. It is an illuminating experience to try the same query on each system, and trace the differences in retrieved items to the underlying search mechanisms.

A demonstration version of the ConSearch commercial product provides an interesting variation: it searches The Virtual Yellow Pages®, a WWW directory created and administered by Interactive Marketing Services, Inc. The ConSearch software has a thesaurus that can automatically augment queries with common synonyms and related terms. This capability is not common in commercial or publicly available IR systems, though the technique has been extensively studies by the IR research community [1].

Subject-specific "digital libraries"

The most common type of "digital library" is an Internet-accessible repository of one or two types of document (technical reports, journal articles, pre-prints, or conference proceedings), specific for a single discipline. The full text of documents are often available, as well as bibliographic records. These libraries build their collections in one of three ways:

- *by harvesting existing documents*. Documents relevant to the subject are located and indexed, with the index alone stored at a central site while the documents themselves remain at their original locations. Examples of this type of organization include UCSTRI [2] and the New Zealand computer science technical report libraries [3].

- *by providing a uniform interface for a limited set of sites*. Participating sites agree to standard formats for document storage, and a central index provides uniform search access for all sites. The WATERS system, for example, indexes technical reports from 14 universities ([4], [5]), and the DIENST system provides an interface to the computer science technical report repositories for an additional 14 leading universities (formally known as the CSTR library) [6].

- *by indexing donated materials*. The digital library is a central location for a discipline, to which authors send descriptions of their papers (ie, title, authors, institutional affiliations, etc) and either a URL location or (less commonly) an electronic form of the paper itself. The physics E-PRINT ARCHIVE is a primary example of this type, and its architecture has been successfully transferred to a number of other disciplines ([7], [8]).

As with the general WWW search engines, discussion of the underlying mechanisms needed to efficiently create, store and search indexes is of interest for an IR course - or, indeed, as an applications case study for a class on searching and the implementation details of search engines. The New Zealand system in particular is supported by substantial documentation that details the data structures, compression techniques, and alternative representations for its search engine ([3], [9]).

The method for creating the collection dictates the type of indexing and searching that the digital library can support. Since libraries built on author submissions generally require the authors to catalog their works, these digital libraries generally offer the same types of search that are available in other online bibliographic systems: by author, title, date of publication, etc. "Harvesting" systems must either rely on simple keyword searches, or extract document summaries (which often vary in quality). It is a useful exercise to explore the effectiveness of these different search capabilities and relate them to the way that the digital library was constructed.

A serendipitous advantage of these digital libraries is that many of the existing collections focus on "grey literature" in the form of pre-prints and technical reports, which is notoriously difficult to search and retrieve in more traditional bibliographic systems. These collections are invaluable resources for senior and graduate students in that they provide fast, inexpensive access to unpublished or newly-published research. Searching exercises can be tailored to individual student interests, and can be used to support work in other graduate courses.

Finally, these digital libraries provide excellent platforms for information science research (particularly bibliometric analysis), and can be useful as a source for exercises or projects in an information science course. These systems provide search engines capable of supporting conventional studies (eg, document obsolescence, trend analysis, etc.). Interestingly, these electronic text archives also provide opportunities for new types of studies: generally the full text of documents are available for analysis, giving a finer grain of insight than abstract-only online databases; and document "usage" can be measured directly by recording user accesses, rather than studied indirectly through document references [10].

Bibliographic databases

A wide variety of bibliographic databases are available over the Internet, including library catalogs and subject bibliographies. These more traditional bibliographic sources are useful in, for example, providing experience with a large number of library OPACS, so that students may compare and contrast a number of vendor software systems; introducing students to common bibliographic formats such as BibTex; and allowing students to compare formally developed bibliographic systems (such as OPACS) with "homegrown" bibliographic databases created by user communities.

A popular commercial bibliographic database that is searchable over the Internet is Carl's UnCover, which indexes 15,000 English-language periodicals. A user can also set up a "profile", or lists of terms describing documents of interest, and the user is automatically emailed notification of new documents matching the profile. Unfortunately, UnCover indexes only the *table of contents page* for periodicals: generally no abstract in available for searching, and users cannot do fielded searches for authors. Novice users generally misunderstand this point, and assume that the "full text" searches they conduct are indeed on the entirety of the document - and therefore overestimate their ability to retrieve documents of interest with simple one- or two-term searches. In the classroom, this system is useful in illustrating the additional searching power that is gained by augmenting indexes with abstracts or other document keywords.

Filtering programs

The IR engines discussed above process queries against relatively static document collections. Filtering systems are essentially electronic clipping services, designed to extract documents of interest to a user from a dynamic stream of incoming items. Two example systems are SIFT, a research prototype developed by Stanford that processes a user "profile" against daily postings on the Usenet News ([[11], [12]); and HeadsUp, a commercial program that filters information from a wide variety of sources (HeadsUp provides free trial subscriptions). Filtering systems are particularly interesting for classroom use in that the terminology used in the documents tends to change rapidly, forcing students to update profiles relatively frequently to maintain reasonable performance for their queries. While this type of jargon or usage shift occurs in bibliographic databases as well, it generally emerges over a much longer time span and is much more difficult to practically demonstrate in a hands-on assignment.

HeadsUp and SIFT are based on content-based filtering: a user's likely interest in a document is calculated from features directly extracted from the document (usually keywords). Collaborative filtering techniques additionally incorporate other people's opinions in the query/document matching process, to refine the matching or relevance ranking by adding subjective information on a document's quality, or its suitability for various audience types. Two collaborative systems from MIT's Multimedia lab are HOMR, a music recommendation system that bases its suggestions on analysis of similarities between user listening tastes, and Webhound, which provides advice in WWW pages likely to be of interest to a given user (again based on previous user opinions).

Classification systems for Internet resources

The Internet is a notoriously "loose" information resource: most documents are presented without formal cataloging, and are accessible only through document linkages or keyword searchers. A few, informally developed, classification schemes have emerged, the most notable among them being the Yahoo Internet Directory. The Yahoo scheme is based on subdivisions of 14 broad initial categories, and WWW page classification is done manually (by Yahoo editors and/or WWW page submittors). Other "homegrown" classification schemes include: the WWW Consortium's category subtree; the McKinley Internet Directory's Magellan subject directory of over one million sites (over 300,000 of which are reviewed and rated); or the Berkeley Public Library's subject guide (with 38 initial subject headings and no further topic subdivisions).

Several experimental indices to WWW pages are based on previously existing classification schemes: the WWW Consortium also maintains a Library of Congress Subject Headings classification of the pages in its index; the NORDINFO project is experimenting with automatically indexing WWW pages by the Universal Decimal Classification (UDC) system; and WebWise, based at the University of Wolverhampton, classifies selected sites (primarily in the UK) according to the Dewey Decimal system.

From a cataloging and classification point of view, it is interesting to investigate these systems from the point of view of the applicability of these classification schemes to the very general WWW page collections. The "homegrown" classification schemes have been developed mainly on the fly, by Internet users and developers rather than by trained catalogers; not surprisingly, these classification schemes are sometimes more difficult to traverse than standard classification systems such as Dewey and UDC. Clifford Stoll provides an entertaining and ascerbic discussion of the shortcomings of amateur Internet classification schemes [13]. However, the standard systems were not developed with WWW pages in mind – so evaluating their suitability for this new type of document is an interesting exercise. Finally, the NORDINFO project is attempting to provide automated WWW document classification; the issues of classification quality in this system and more technical details of the classification algorithms can be explored.

Image databases

Image databases present a different set of problems than text collections. Most image storage and retrieval programs are based on text description of images - an approach with obvious problems, since spatial relations, textures, and colors are notoriously difficult to adequately and accurately catalog in words. QBIC is a demo version of an IBM product that permits users to construct queries directly based on image characteristics. The query types available in the Internet-accessible version of QBIC are simple - color percentages and distributions - but illustrate the flexibility and effectiveness of constructing queries based on image processing-type information.

Demonstrations and tutorials are also available for Geographic Information System (GIS) programs, one of the most widely used types of spatial/visual retrieval and querying systems. While these resources do not allow users to create a GIS from their own data and generally restrict users to pre-determined queries, the WWW software does give a good flavor of the types of queries and representations supported by a GIS. Students can, for example, complete a detailed ARC/INFO tutorial, which includes an option for creating 3D visualizations of the Holy Land; use NAISmap to view and manipulate National Atlas spatial data to construct a personalized map of Canada; visualize animal and plant species distribution in Australia; or use MAPPER to interactively define a region of the world and then link that region with informational queries (ie, temperature, city location) drawn from a relational database.

Benefits and drawbacks for classroom use

Our experiences in using these systems in the classroom have generally been good. The IR programs discussed above have a number of advantages in illustrating retrieval techniques and algorithms:

- They are freely available, either as ftp-able source code or as interactive programs accessible through the WWW. Most are full-fledged applications, and the remainder provide useful and useable demonstrations.

- Source code is available for many programs. These programs can be set up for experimentation at the user's own site (avoiding problems with access times, and permitting the students to work with any special document collections available locally). More technically oriented students can explore implementation techniques by directly examining the source code.

- The programs are applied to a variety of realistic document collections, many of which are of direct interest as reference resources (eg, technical report collections).

- Detailed documentation is available for most programs, with the exception of commercial products. Many systems are extensively described in technical reports or journal/conference articles, which are themselves often available over the Internet.

However, there are a number of disadvantages associated with using these programs as well - generally the same disadvantages found with using free software or demonstration versions of commercial software in courses [14]:

- Documentation is of variable quality. Programs developed by university research groups tend to be documented by their published papers. Unfortunately, when the system changes the alterations are sometimes not described in similar detail (or in some cases, at all!).

- Research prototypes are not supported as well as commercial products; document collections may be updated at erratic intervals, systems may be temporarily unavailable, and query processing times may be inexplicably lengthened. The key to success in using these systems for course assignments is to allow sufficient time for students to access the systems, and to emphasize to the students that they cannot put these assignments off to the last minute (when the systems may be down and inaccessible).

- IR systems accessed over the WWW suffer from general Internet problems - occasional difficulties in connecting and slow feeds. These delays can be frustrating for students, particularly when they are used to the stable access offered by commercial retrieval systems or local OPACS and CD-ROMs.

Conclusions

The current range of Internet-accessible IR programs present an exciting opportunity to enrich the curriculum of information retrieval courses. Students can be exposed to the basics of IR algorithms and data storage and retrieval mechanisms through lectures and their textbook, but these newly-available programs also give them an opportunity to gain experience with real application and document collections. Assignments can demonstrate the effects of different search term weighting schemes, relevance ranking techniques, stemming algorithms, etc., as well as the computational bases for these pieces of software.

As an additional benefit, students are exposed to a variety of documents useful to their field of specialization - technical reports, WWW pages, Usenet News discussions, etc. Students can learn about effective techniques for document and literature searches as they explore a rich variety of knowledge sources.

References

1 Salton, G., and McGill, M.J. *Introduction to modern information retrieval*. New York: McGraw-Hill, 1983.

2 Van Heyningen, M. The Unified Computer Science Technical Report Index: lessons in indexing diverse resources. *Proceedings of the Second International WWW Conference*, (1994).
<URL:http:///www.cs.indiana.edu/ucstri/paper/paper.html>

3 Witten, I.H., Cunningham, S.J., Vallabh, M., and Bell, T.C. A New Zealand digital library for computer science research. *Proceedings of Digital Libraries '95* (1995) 25-30.

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

5 French, J., Fox, E., Maly, K., and Selman, A. Wide area technical report service: technical reports online. *Communications of the ACM* 38(4) (April 1995) 45.

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

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

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

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

10 Cunningham, S.J., Empson, N., and Kamau, R. Bibliomania: what can we learn from the research literature? *Proceedings of the New Zealand Computer Society Conference*, 1995.

11 Yan, Tak W., and Garcia-Molina, Hector. Distributed Selective Dissemination of Information. Proceedings of the Third International COnference on Parallel and Distributed Information Systems, Austin TX, 1994. Also available at <URL:ftp://db.stanford.edu/pub/yan/1994/dsdi.ps>

12 Yan, Tak W., and Garcia-Molina, Hector. SIFT -- A Tool for Wide-Area Information Dissemination. Proceedings of USENIX '95, New Orleans LA, 1995. Also available at <URL:ftp://db.stanford.edu/pub/yan/1994/sift.ps>.

13 Stoll, Clifford. *Silicon snake oil : second thoughts on the information highway*. New York : Doubleday, 1995. Chapter 11 ("Wherein the author considers the future of the library, the myth of free information, and a novel way to heat bathwater"), pp. 173-214.

14 Myers, W. Instructional uses of demonstration disks. *SIGCSE 'Bulletin* 26(1) (1994) 76-79.

Appendix: IR system location information

World Wide Web search engines:

InfoSeek	http://www2.infoseek.com
Lycos	http://lycos.cs.cmu.edu
Open Text	http://opentext.uunet.ca:8080/omw.html
WebCrawler	http://webcrawler.com/
ConSearch	http://www.imsworld.com/miti/ (Virtual Yellow Pages WWW directory)

Subject-specific digital libraries:

NZ CS technical report library	http://www.cs.waikato.ac.nz/~nzdl/
UCSTRI	http://www.cs.indiana.edu/cstr/search
WATERS	http://www.cs.odu.edu/WATERS/WATERS-GS.html
CSTR	http://cs-tr.cs.cornell.edu/Info/cstr.html
physics e-print archive	http://xxx.lanl.gov/

NASA list of technical report digital libraries:
http://www.larc.nasa.gov/org/library/abs-tr.html

Bibliographic databases:

UnCover (commercial database and access to library online catalogs)	http://www.carl.org/uncover/brochure.html
Computer science bibliography with search engine (primarily BibTex)	http://liinwww.ira.uka.de/bibliography/index.html
list of library OPACs with WWW interfaces	http://www.lib.ncsu.edu/staff/morgan/alcuin/wwwed-catalogs.html

Information filtering software:

HOMR music recommendation system	http://rg.media.mit.edu:80/projects/homr/
Webhound	http://webhound.www.media.mit.edu/projects/webhound/
SIFT	http://sift.stanford.edu/

Information filtering resources list:
http://www.glue.umd.edu/enee/medlab/filter/filter.html

Classification systems applied to WWW documents:

Yahoo Internet Directory	http://home.netscape.com/home/internet-directory.html>
Magellan (McKinley Internet Directory)	http://www.mckinley.com
Berkely Public Library Index to the Internet	http://www.ci.berkeley.ca.us:80/bpl/bkmk/index.html
WWW Consortium LCSH index	http://www.w3.org/hypertext/DataSources/bySubject/LibraryOfCongress.html
WWW Consortium category subtree	http://www.w3.org/hypertext/DataSources/bySubject/Overview.html
NORDINFO project UDC classification	http://www.ub2.lu.se/autoclass.html

image database and GIS software:

QBIC	http://wwwqbic.almaden.ibm.com/~qbic/qbic.html
ARC/INFO tutorial	http://boris.qub.ac.uk/shane/arc/ARChome.html
NAISmap map of Canada	http://ellesmere.ccm.emr.ca/cgi-bin/naismap/naismap1?e
Australian map interface	http://kaos.erin.gov.au/cgi-bin/spatial_interface
MAPPER	http://aurora.esri.utah.edu/map-cgi-bin/mapper.py

GIS resource lists:
http://www.geo.ed.ac.uk/home/giswww.html
http://www.blm.gov/gis/nsdi.html