Information Retrieval Tools and Techniques

A. Accomazzi

Smithsonian Astrophysical Observatory, Cambridge, MA, USA

F. Murtagh

Space Telescope – European Coordinating Facility, European Southern Observatory, Garching near Munich, Germany. (Affiliated to Astrophysics Division, Space Science Department, European Space Agency.)

B.F. Rasmussen

Space Telescope – European Coordinating Facility, European Southern Observatory, Garching near Munich, Germany

Abstract

Retrieval of data and information is something which every librarian, scientist and technologist does many times over in each working day. Much progress has been made in recent years, and we overview search engines and distributed toolsets. However there are still some problems to be overcome before heterogeneous data can be integrated for the user in a seamless way, chiefly in regard to seamless integration of image data into current practices. We present a short state-of-the-art overview of the outstanding achievements of recent years; and of some of the more challenging, and potentially fruitful, open issues.

1 Introduction

Database management systems (DBMSs) are concerned with retrieval based on exact and partial match searching. Information retrieval (IR), on the other hand, is concerned with best match searching. For DBMSs, the problem becomes one of structuring the data, and providing user views on the data. For IR, indexing is a necessary first step, followed by querying, which supports greater or lesser expressiveness. See [6,17,18] for introductory readings.

A topical area of IR research and development work relates to multimedia and multitype data. Astronomy offers lush pastures for such work, in particular as regards text and image data.

We begin here with a review of astronomical textual and bibliographic tools and techniques (sections 2 and 3). This will be broadened (section 4) to discuss what is now being done to integrate text and image data. Section 5 addresses near-future issues in regard to treatment of images: to use programming language parlance, should images be compiled, before returning their informational result, or instead interpreted? As with programming languages, both approaches have an important role to play.

2 Free Text Retrieval, Supporting Phrases, Using lq-text

This section reviews a simple text search utility.

Lq-text (authored by L. Quin, available by anonymous ftp from ftp.cs.toronto.edu), is a text information retrieval utility, using indexing, which supports compound terms and keyword-in-context. (Lewis and Jones, [8], review available technologies for handling compound terms, which they see as aiding in semantically-oriented retrieval.) It does not support the Z39.50 or related protocols, and is not immediately interfaceable to WAIS (to be discussed in section 3).

Advantages of *lq-text* include: compound term retrieval is supported, including phrases which are broken by ends of lines; dashes are ignored ("Lyman-alpha", "Lyman alpha"); and in the default mode, case is ignored. Disadvantages include no specific catering for astronomical semantics (in the default setting, words are between 3 and 18 characters in length, and thus "UV" is ignored); boolean queries are supported but awkwardly, by ranking and intersecting; and strings which consist entirely of numeric data, or start with numeric characters, are excluded. It is however a tool which can be used as the basis of a larger retrieval system.

The following Unix line-mode example (the system prompt is `command>`) shows compound term support.

```
command> lqphrase -v "IRAS galaxies"
Word IRAS --> Iras, 32 matches
Word galaxies --> galaxies, 1981 matches
3       38      1   3   hstprop3913
```

Although there were 32 matches with the word "IRAS", and 1981 with "galaxies", only in document "hstprop3913" did these come together as a phrase. We are not interested here in the remaining numeric information returned, relating to offsets in the text of the word found. A KWIC (keyword in context) option follows.

```
command> lqkwik 'lqphrase "IRAS galaxies"'
91) that ultraluminous   IRAS galaxies are      : hstprop3913
```

Another support option is to show more of the context of what has been found. A screen with up to 6 (default) lines before and after the compound term or word, for each hit, is presented using command "lqshow".

Multiple phrase support ("IRAS starburst galaxies" "spectral energy distributions" "massive galaxies" "elliptical galaxies" "protogalaxies") is also supported. We get information on the number of hits associated with the individual words in the phrase.

3 WAIS and Friends

This section reviews network-based retrieval of text and (increasingly) data in other forms.

The modern computing system is constructed from many local and remote machines with the result that the client-server computing model has become a central concept in such systems. In IR, client-server systems are widely used, and some important recent and current developments in regard to tools and techniques are now looked at.

WAIS (Wide Area Information Servers) is a widely used client-server information retrieval system. WAIS originated in a joint research project between Thinking Machines Inc. (recently experiencing financial difficulties), Apple Inc., Dow Jones & Company, and KPMG Peat Marwick, and was first released in 1991. The formation of WAIS Inc. by WAIS principal B. Kahle led to support of a freely-available version being assumed by CNIDR (Clearinghouse for Networked Information Discovery and Retrieval, located at MCNC, Research Triangle Park, North Carolina).

Z39.50 is an information search and retrieval protocol. Z39.50-88 (Version 1) is used by WAIS. This standard is ANSI/NISO (American National Standards Institute/ National Information Standards Organization) Z39.50, which was successfully balloted in 1988. WAIS as developed by CNIDR, freeWAIS, changed its name in 1994 to ZDist, and later imported into Isite, to signal the move from support of Z39.50-88 to Z39.50-92.

Z39.50-92 (Version 2) is widely used in libraries and information organizations. This is an ANSI/NISO standard. This version involved alignment with the ISO (International Organization for Standardization) SR (Search and Retrieval) protocol (ISO 10162/10163), among other changes. Z39.50-88 and Z39.50-92 are not compatible. Support for Z39.50-92 is provided by Isite (see below).

Z39.50 Version 3 was balloted in recent months by the Z39.50 Implementors Group (ZIG), which works closely with the standard's maintenance agency, the Library of Congress. To be informed about this group's activities, subscribe to list `z3950iw` at address listserv@nervm.nerdc.ufl.edu.

Here are a few widely-used WAIS implementations:

- WAIS release 8 beta 5 minor release 1 (WAIS 8b51), released in May 1992 by Thinking Machines Corp. (ftp to think.com). An Indiana University version (IUBio; ftp to ftp.bio.indiana.edu) supported boolean search.
- freeWAIS version 0.5 (beta), the current version of this series, from CNIDR (support for Isite has taken over from freeWAIS). Boolean search and support for multitype files are available. CNIDR can be accessed at URL ftp://ftp.cnidr.org/pub/NIDR.tools.
- Isite, from CNIDR, supports Z39.50-1992 (version 2). See http://vinca.cnidr.org/software/Isite/Isite.html. Recent updates have included fielded searching and support for SGML tag parsing.
- freeWAIS-sf, from the University of Dortmund, was based on freeWAIS. Further information may be obtained at: http://ls6-www.informatik.uni-dortmund.de/freeWAIS-sf/. Supported features include: full boolean search; text, date and numeric field structures; configurable headlines; 8-bit character support; stemming and phonetic coding (the latter using the `soundex` and `phonix` approaches: see [14]).

Various WAIS–WWW gateways or scripts are available. For further information refer to a number of items accessible from http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/D2-complex.html (in turn accessible under "Manual" in the NCSA Mosaic standard help pull-down menu).

Other similar tools can also be set up as scripts. As an example, Glimpse (GLobal IMPlicit SEarch) is an indexing and querying system which allows files in possibly nested directories to be searched through. It is based on a storage-efficient index, and on `agrep` which is an extension ("approximate grep") of the well-known Unix command, which supports approximate matching (misspellings, etc.). GlimpseHTTP is an automatic HTML indexer. Glimpse is available from the University of Arizona, Department of Computer Science (for source code and documentation, refer to ftp://cs.arizona.edu/glimpse/). We note that freeWAIS-sf has also been updated (in a third-party patch) to handle proximity-based queries.

Another interesting development relates to a Spatial WAIS standard. This is under the guidance of the FGDC (Federal Geospatial Data Clearinghouse). The aim is to incorporate a spatial metadata standard into WAIS. Although oriented towards GIS (geographic information systems), the inspirational and technical relevance for astronomy is clear. Support is available for such spatial extensions as: "at a point location"; "in a bounding box"; and "in a region". In the indexing phase, some additional index files are created to support these spatial queries. Discussion of converging such spatial prototypes into the Z39.50 Version 3 standard are conducted on list zmap, subscribable to at address listserv@vinca.cnidr.org. Discussions in other communities – e.g. British GIS – are ongoing: a useful summarization of spatial data standards issues can be found in [9]).

One final pointer here is to STAS, the scientific and technical attribute and element set, which seeks to define standard identifiers for referring to searchable and retrievable fields within scientific, technical and related data collections. These standards use the Z39.50 protocol. Other than CNIDR, sponsors include the American Chemical Society (ACS, more strictly their Chemical Abstracts Service, CAS) and commercial information providers such as Dialog Information Services and FIZ (Fachinformationszentrum, Karlsruhe).

4 Integration of Text and Image Databases

In this section, we describe recently-installed search trajectories, including support for free text, preview images, full-size images, image ancillary information, and bibliographic data.

Technical support for queries consisting of text chunks (observing proposal abstracts, sets of keywords, bibliographic abstracts, etc.) is straightforward with WAIS, since each term is considered by default to have a boolean OR connective with the following term. This allows support for queries such as: give me all abstracts which are similar to a given abstract.

Clustering of text chunks can be carried out along these lines. Murtagh [11] detected terms (based on the IAU Thesaurus, thereby providing a controlled vocabulary) in around 1500 HST observing proposal abstracts, and clustered them on the basis of their shared term-set. A Kohonen self-organizing map approach was used.

Handling of multitype data (e.g. images and accompanying texts) may be achieved by multitype support in the search engine. Thus freeWAIS-sf can return a reference to the text providing hits, as well as accompanying images, and either can be looked at independently by the user.

An alternative is to structure text which is returned in response to hits such that links to accompanying images are embedded. (In fact the images themselves can be embedded also: [12]).

A search trajectory, embodying free text capability, image quickview, and bibliographic literature search, was set up along the following lines for the HST image database (see http://www.eso.org/hst-prop-abs-search.html):

1: Using free text on proposal abstracts, and/or on fields associated with principal investigator (PI) or proposal identifier number, the proposal title can be obtained, together with its approximate matching score. ($1 \longrightarrow 2$).
2: From the proposal title, by clicking, one gets further proposal information: PI details, observing cycle number, full abstract, and list of exposures. ($2 \longrightarrow 3$).
3: From the list of exposures, one obtains the associated information on the instrument, a link to background information on the instrument, and an indication of availability of preview images (i.e. compressed versions of the real images). ($3 \longrightarrow 4$).
4: From the preview images, one can view these, or mark them for batch retrieval of the real images.
5: From the abstract, or the authors, or the keywords, one can proceed to a search of all relevant entries in the ADS Bibliographic Service. ($2 \longrightarrow 6$).
6: From the ADS Bibliographic Service, one can find other similar abstracts; or one can obtain copies of full papers in some cases. ($6 \longrightarrow 6$ and $6 \longrightarrow 7$).
7: The search may be terminated with published papers, and data tables.

One sees, here, that most linkages are technically feasible. Enabling technologies for these browsing paths, other than freeWAIS-sf, include the WDB Web-to-SQL database interface tool ([16]); and the ADS Bibliographic Service search engine ([1]). Browsing trajectories, as illustrated here, are limited far less than in the past by the particular data-type (images, texts). To a greater extent than heretofore, the user is allowed to express their information needs in something which approaches natural language. We have even seen with `soundex` and `phonix` (section 3 above) that vocal querying is not far off.

Two issues are still unresolved: how best to organize such browse and search trajectories (an issue which we will not pursue further, here); and what about direct querying of image contents?

5 Adding Intelligence to Images

In this section, we describe ongoing work in content-based image IR; and in making images active and interactive.

One approach to content-based image retrieval is through image indexing, i.e. through creation of an object catalog or inventory. For building up an inventory of the image's contents, traditional approaches to object inventory have used thresholding. A more thorough approach to adaptive thresholding is preferable, which relativizes background and incorporates a vision model relating to the type of object which is sought as a priority. Since a vision model is based on multiresolution analysis, and mathematical morphological operations. Bijaoui and Rué [4] present a comprehensive view of work in this direction.

Multiresolution analysis is motivated by the fact that the human visual system deals with visual scenes at differing resolution scales. It handles these resolution scales simultaneously. Using wavelet transforms, or other approaches ([19]), the first phase is to arrive at a set of resolution scales representing different levels of information in the original image. Multiresolution transforms have been shown to be an excellent basis for noise suppression in the image, and for image compression. The multiresolution support ([19]) is a data-structure which may be derived from an image: it is the significant (or interesting) part of the image, at a given resolution level. The multiresolution support is a multilevel boolean image, where contiguous sets of 1-valued pixels demarcate astronomical images of interest. A priori knowledge of objects or detector defects may be incorporated by mathematical morphology operations on the multiresolution support. An inventory is built up of the objects found.

Insofar as object templates are imposed on the given image in such an approach (cf. specification of the multiresolution transform used, or incorporation of a priori information through erosion and dilation operations, etc.), and insofar as idealized astronomical objects are output at the end of the processing (in the form of inventory tables), what we are really doing here is transforming the input image into generalized icons. The icons are (if successful) our desired astronomical objects (point sources, spiral galaxies, etc.).

An alternative approach to image understanding is to start with the image itself, and make its indexation interactive. This is in contrast to starting with our expert judgement about what it contains, which is then encapsulated in a multiresolution analysis system or some other alternative analysis system. Making the image's indexes active is an objective which is currently being addressed in Strasbourg Observatory's ALADIN project ([13]). Clicking on the interactive atlas will provide links to various sources of information (astronomical catalogs, bibliographic data).

Chang and Hsu [5] already claim to see beyond image modeling (generalized icons) and active indexes, and to perceive the contours of smart images. These are images with associated knowledge structures (which can be thought of as generalizing current practice of associating descriptors and audit trails with images). Images would have associated context-dependent active methods for display (cf. pyramidal data structures, preview functionality, or iconization as discussed above); transmission (cf. progressive transmission using http and other protocols); hyperlinking (cf. ALADIN); data structures for support of processing algorithms; and interrelationships with other images in the same and in other image databases.

Intelligent search agents, which operate network-wide, are described in [3]. Active agents are also central in current computing research directions, in particular as regards biological computing ([7]). What is described as "smart images" in [5] are active vision agents, which play their role in future computing environments.

6 Conclusions

How best to link not just data, but information, in whatever form it presents itself, has been at issue in this article. Major problems have been domesticated in recent years, e.g. handling distributed, multiformat information. We have pinpointed some open issues, chiefly in regard to content-based image retrieval.

Such a problem comes within the scope of a 3-year international consortium funded by the European Science Foundation to study "Converging Computing Methodologies in Astronomy" and to provide a policy recommendation at the end of this period. Further information on the consortium can be found at URL http://www.eso.org/conv-comp.html.

References

[1] A. Accomazzi, G. Eichhorn, C.S. Grant, S.S. Murray and M.J. Kurtz, "The Astrophysics Data System Abstract and Article Services", *Vistas in Astronomy* (special issue, proceedings of WAW Conference, April 1995), 1995, in press.

[2] A. Accomazzi, "Notes on the use of search engines at ESO", 1995, http://www.eso.org/wais-documentation/fwsf-notes-aaccomaz-apr95.html

[3] H.-M. Adorf, "Resource discovery on the Internet", this meeting, 1995.

[4] A. Bijaoui and F. Rué, "A multiscale vision model adapted to astronomical images", 61 pp., 1994, submitted.

[5] S.-K. Chang and A. Hsu, "Image information systems: where do we go from here?", *Handbook of Pattern Recognition and Computer Vision*, eds. C.H. Chen, L.F. Pau and P.S.P. Wang, World Scientific, Singapore, 941–965, 1992.

[6] W.B. Frakes and R. Baeza-Yates, Eds., *Information Retrieval: Data Structures and Algorithms*, Prentice-Hall, 1992.

[7] A. Kay, Apple Computer Inc., "When can we take the training wheels off?", plenary presentation, Third International WWW Conference, Darmstadt'95, 1995.

[8] D.D. Lewis and K.S. Jones, "Natural language processing for information retrieval", preprint, AT&T Bell Laboratories and Computer Laboratory, University of Cambridge, 29 pp., 1995.

[9] D. Medyckyj-Scott, C. Monckton and P. Burnhill, "Progress towards standards for spatial metadata", 14 pp. draft document (GENIE Project, Dept. Comp. Sci., Loughborough Univ. Tech.; and RRL Scotland and Data Library, Univ. Edinburgh), 1995.

[10] J.J. Michael and M. Hinnebusch, *From A to Z30.50: A Networking Primer*, Mecklermedia, Westport CT, 1995.

[11] F. Murtagh, "Free text information retrieval: an assessment of publicly available Unix-based systems", technical note, 12 pp., 1994.

[12] F. Murtagh, W. Zeilinger, J.-L. Starck and A. Bijaoui, "Object detection using multiresolution analysis", in D. Shaw, H. Payne and J. Hayes, Eds., *Astronomical Data Analysis Software and Systems IV*, Astron. Soc. Pac., 1994, in press.

[13] Ph. Paillou, F. Bonnarel, F. Ochsenbein and M. Crézé, "ALADIN: Atlas Interactif du Ciel Profond", CDS Observatoire Astronomique de Strasbourg, report, 79 pp, Version 1.3, 1993.

[14] U. Pfeifer, T. Poersch and N. Fuhr, "Searching proper names in databases", to appear in *HIM95: Proc. Conf. Hypertext-Information Retrieval-Multimedia*. Available at: ftp://ls6-www.informatik.uni-dortmund.de/pub/doc/reports/95/Pfeifer-etal-95a.html

[15] U. Pfeifer, N. Fuhr and T. Huynh, "Searching structured documents with the enhanced retrieval functionality of freeWAIS-sf and SFgate", *Computer Networks and ISDN Systems*, 27, 1027–1036, 1995.

[16] B.F. Rasmussen, WDB software and documentation, http://arch-http.hq.eso.org/bfrasmus/wdb/, 1995.

[17] G. Salton, "Developments in automatic text retrieval", *Science*, 253, 974–1015, 1991.

[18] G. Salton, J. Allan, C. Buckley and A. Singhai, "Automatic analysis, theme generation, and summarization of machine-readable texts", *Science*, 264, 1421–1426, 1994.

[19] J.-L. Starck, F. Murtagh and A. Bijaoui, "Multiresolution support applied to image filtering and restoration", *Computer Vision, Graphics, and Image Processing – Graphical Models and Image Processing*, 1995, in press.