Resource Search and Discovery

Gary Marchionini
College of Library and Information Services
University of Maryland

Electronic technology has begun to change what information is available and how that information is located and used. There are already a large number of electronic projects in the humanities (see Getty, 1994). These changes are first related to remote access: Instead of traveling to the sources of information, scholars use technology to bring information to them. One important consequence of remote access is the broadening of access to students and other novices who would not or could not bear the time and financial costs to travel to libraries, museums, and research institutes, and who might not know what to look for once they arrived. Secondly, electronic technology brings new genres of information that provide new challenges for search and discovery (e.g., multimedia, interactive ephemera, etc.). The traditional problems humanists have found in documenting and locating non-textual materials are exacerbated by electronic technology. Thirdly, change is due to electronic tools and the strategies made possible by electronic representations. The emphasis here is on tools and strategies for resource search and discovery, although I will argue that we will continue to see closer integration with tools and strategies for creating, using, and communicating information. This implies that creators who choose to become more closely involved with consumers must take more responsibility for documenting and directly placing their work.

In archives, libraries, and museums, search and discovery are facilitated by finding aides, catalogs, and guides that organize the information space for information seekers. It is evident that similar devices are appearing for electronic resources as well. An ongoing research challenge is to discover appropriate representations for information and new search and discovery tools and strategies that leverage the computational medium.

Search implies an effort to locate a known object; the information seeker has in mind specific characteristics or properties of the object and these characteristics are used to specify and guide search activity. **Discovery** implies an effort to explore some promising space for underspecified or unknown objects; the information seeker has in mind general characteristics or properties that outline an information space in which perceptual and cognitive powers are leveraged to examine candidate objects (elsewhere I have distinguished search and discovery as analytical and browsing information seeking strategies respectively, Marchionini, 1995). In general, discovery emphasizes the location of the promising space (a collection or resource (e.g., CNI, in preparation). Electronic technology provides new tools for each of these classes of strategies and also blurs the traditional boundaries between them.

State of the Art

Scholarly search and discovery have long depended on mappings between conceptual space and physical locations: Classification systems organize information objects, thesauri map these organizations onto word labels, and catalogs provide pointers from labels to physical objects. Traditionally, there have been clear demarcations between the n-ary information objects such as indexes and catalogs, and primary information objects such as books and physical artifacts. The

Internet includes n-ary and primary information objects and todayÕs interfaces make little distinction between these representations; effectively blurring these boundaries. Thus, electronic technology influences information seeking by changing both the traditional tools that support search and the strategies we use for information seeking. Any attempts to develop cataloging schemes for Internet resources must take into account these differences as well as address the difficulty of documenting dynamic and ephemeral information objects such as ftp and web sites. Moreover, it is certainly too soon and likely wrong to aim at developing collection development policies and a master catalog for the Internet as a whole. Nonetheless, specific digital libraries and resource collections have begun to take advantage of information retrieval and information seeking research to make information more easily and readily available.

Search. Information retrieval research has yielded several approaches to the problem of matching queries to documents and object surrogates. Traditionally, these approaches have been applied to specific collections of documents (one set of resources) rather than across many different collections. The most basic advantage of text in electronic form is the ability to do string search--to locate all occurrences of a string of characters in a text or corpus. Although many algorithms are used to support string search, inverted file indexes are used in most large-scale systems to support free-text searching. Building upon string search techniques, scholars are able to develop concordances (e.g., the Dead Sea Scrolls) and explore word usage frequencies across authors or works (e.g., Thesaurus Linguae Graecae with Pandora). Although many of these efforts are currently restricted to stand-alone, proprietary collections, some are available through the Internet. There has been little progress in indexing non-textual materials although scene changes and color patterns have been used to augment video and graphical databases. Most non-textual objects are located through textual descriptions or linear scanning.

Another major development in search is the ability to rank documents according to one of many statistical or probabilistic algorithms. These algorithms use word or phrase frequency data to match queries with documents and rank results accordingly. Although computationally intensive, todayÕs computers are able to manage representations of documents as n-dimensional vectors and compute similarity measures for documents and queries in n-dimensional space. These approaches have gained commercial appeal (e.g., DialogÕs Target and Lexis/Nexis Freestyle) and many Internet resources are now using statistical or probabilisitc search engines on their servers (e.g., several WAIS-based services are available; the Library of Congress Thomas system uses the Inquiry search engine). In most cases these approaches provide ÒkeywordÓ access (based on all words in the corpus except some small set of common words) rather than ÒsubjectÓ access (based on a carefully constructed controlled vocabulary used by indexers to describe the content of the object). Although ranked retrieval offers good advantages to novice searchers and a viable alternative to Boolean-based search for experienced searchers, we are a long way from providing all and only relevant information objects to information seekers who pose word-based queries (see the Center for Intelligent Information Retrieval web site for information on Inquiry http://ciir.cs.umass.edu/).

A third set of approaches to search leverages the logic of discourse or substantial knowledge bases to contextualize queries or possibly modify them. For example, the Perseus system (Crane, 1992) includes a morphological analyzer that goes beyond string search to provide variant forms for Greek words. Some linguists aim to develop generic grammars that represent the domain of

logical statements possible and parsing routines that map natural language queries and documents onto the grammar. Other researchers have developed schemes for taking advantage of meta knowledge provided by authors or publishing specialists. For example, the Text Encoding Initiative (see Hockey paper) promotes the use of SGML coding in scholarly texts so that information seekers can use these codes for locating and analyzing texts. Another line of research aims to develop thesauri (e.g., the Art & Architecture Thesaurus) that provide controlled entry points for information seekers as they formulate queries or that are applied automatically to modify or expand queries during the retrieval process. Proficient searchers can certainly use a thesaurus to good advantage but automatic query expansion based on a thesaurus has not generally yielded improved search results (e.g., Jones, Gatford, Rugg, Hancock-Beaulieu, Robertson, Secker, & Walker, 1995; Voorhees, 1994).

A fourth class of research aims to develop filtering systems that automatically route potentially relevant information to scholars. Search depends on specification of the sought object and filtering depends on specification of the user. Traditional selective dissemination of information services have long been provided to scholars by libraries that devote human effort to scan information services according to institutional and individual interest profiles. Online services offer users the opportunity to define interest profiles (usually word based) and then alert the user when information objects arrive that fit the profile (e.g., document delivery services such as UnCover). Different implementations may use any combination of the search algorithms above. In the Internet environment, there are several examples of network news filters that adapt as users provide positive and negative feedback and there are programs of research to develop active agents that roam the network to locate profile-appropriate information and in some cases, cooperate with other software agents (see the Oard web site for a set of pointers to filtering research; http://www.enee.umd.edu/medlab/filter/).

Finally, some research has attempted to automate traditional reference and question answering services. Early efforts used expert system technology to automate selected reference services and todayÕs efforts aim to go beyond the simple frequently asked question (FAQ) services to develop multiple tiers of online reference support (e.g., AckermanÕs Answer Garden for X Windows).

Discovery. Browsing has many attractions for scholars: exploration, contexualization, and serendipity support the discovery of new connections between known ideas and the discovery of new pertinent informational resources. In manual environments, browsing has been done in specific collections (e.g., a section of shelves). Electronic technology in general and the Internet in particular has greatly expanded the universe of browsable material by bringing it to the information seeker at the desktop. Because the Internet connects a multitude of collections (on all topics, in various media, and using different organizational schemes), discovery has become complicated by the need to first limit browsing to a set resources. Developing tools and strategies for identifying resources to browse this thus a primary research challenge.

One form of ÒguidedÓ discovery is exemplified by hypertext systems. Most hypertexts use explicit links denoted by link anchors (buttons, highlighted text) to suggest routes for users to follow. In standalone hypertext systems (i.e., specific collections), users can navigate effectively by following explicit links. Many scholars consider such links to be editorial acts; thus

aggregations of existing materials woven together with hypertext links represent added value derivative works at least and original scholarly interpretations at best. The immense popularity of the World Wide Web (WWW) is based on the ease with which users can follow hypertext links with public domain and easy to use client software often called ÒbrowsersÓ (e.g., Mosaic, Netscape). Hypermedia systems such as Perseus and Piero press the links further by offering implicit or computed links that are made available as the results of queries entered by the user. Electronic texts that provide SGML or other markup codes can also provide on-the-fly link constructions that allow information seekers to follow paths defined by their articulated needs rather than predefined links authors or editors provide. Other approaches include dependencies based on system state (e.g., Petri nets) and scripts that compute links based on user behavior. Even after users have limited their discovery to a set of pertinent resources, personal discipline is required to remain within that set (e.g., todayÕs browsers do not dynamically limit links to those sites contained in a preliminary selection of resources).

Discovery depends both on locating candidate objects and recognizing relationship(s) between those objects and the problem under investigation. The interplay between the perceptual aspects of browsing and the cognitive aspects of reflection and evaluation is best supported by systems that present accurate and well-documented representations (i.e., authors or their agents are explicit about their ÒperspectiveÓ) for objects and allow users rapid and precise control. Direct manipulation interfaces (see Shneiderman paper) best illustrate such interfaces in computing environments. Developments such as the use of thumbnail images as well as text-based descriptions provide new types of surrogates for information objects and support rapid scanning and browsing. Multiple levels of representation for texts are emerging in networked environments as users move from the entire Internet to a subset (possibly ranked) of resource titles to outlines or tables of contents for specific objects to extracts from the objects, to the full representation of the object, and eventually to related objects.

Integration of Search and Discovery. Because electronic environments are blurring demarcations between search and discovery strategies, there are several developments that suggest research directions. First, one way to improve the results of search is to use relevance feedback. Given a set of objects retrieved for a query, users may be able to identify those that are appropriate to the need and those that are not. These judgments are feed back to the system and the original query is either modified or a new query is formulated that combines the original query with the additional information gained through feedback. Relevance feedback illustrates the linkage between search and discovery--a search query serves to identify an intellectual neighborhood for the information to examine (often by browsing) and the results of the examination are used to refine the neighborhood. This process mirrors what information seekers do in manual environments, but the computational tools multiply the number of iterations possible per unit time. Just as rapidly displayed, coordinated still images become moving pictures beyond thresholds of 10-15 frames per second, this quantitative increase may lead to qualitative shifts in search and discovery. One possible avenue of development in this regard is hierarchical (cascading) dynamic query systems.

Another development that improves search and discovery in the Internet is the use of indexing programs called spiders or robots that systematically link to WWW sites, record whether the site has previously been visited, and record basic metadata about each site (sites may also contribute

indexing information voluntarily). These programs have made the WWW somewhat searchable without constraining the browsing features of servers or clients (e.g., Lycos [http://lycos.cs.cmu.edu/] and Yahoo (http://www.yahoo.com/) services allow simple word searching on several million web sites; Yahoo provides a simple classification system for limiting searching). It is important to note that these services do not really represent a catalog of the Internet but rather a listing of Òhome pageÓ words. Additionally, to avoid tying up network resources, spiders do not traverse all links in a site (thus a more substantive image of what the site contains and to what it links is not available). Another system (OpenText [http://www.opentext.com:8080/omw.html]) provides full-text retrieval but also allows searches on SGML tags and supports multilingual searching. Another approach is illustrated by the Harvest project (http://harvest.cs.colorado.edu/) that separates the indexing ÒgatherersÓ from the indexes themselves (ÒbrokersÓ). This allows multiple and customized indexes to be tailored for specific communities.

The most important illustrations of integration are the developments in interactive interfaces that closely couple search, evaluation, and reformulation. Dynamic queries, fisheye views, semantic maps, and other visualization mechanisms are illustrative of such integration. Additionally, the quality of electronic display continues to improve as fonts, backgrounds, color, and resolution continue to offer more accurate representations for paper documents and other information objects (see Shneiderman). One project that tightly couples textual information and graphical information is the Piero project (Lavin, 1992) where relational database entities are linked to 3-D visual database, allowing users to search and discover textually or visually.

Challenges in the Humanities. Although the research and development trends discussed above are applied in all domains, the humanities offers special challenges for search and discovery. First, the humanities celebrate individuality; information resources take many forms and scholars often resist the imposition of standards. These effects are most apparent in word-based searching which is complicated by the opposing concerns of creators who endeavor to find unique and figurative language (whether the language of expression is textual, aural, or visual) and searchers who endeavor to map their needs onto language. Asking authors to use ÒstandardÓ language is ludicrous so it remains for editors, librarians, curators, and other information specialists to create customized indexes and guides to the literature. Furthermore, individuality leads to the creation of many fairly small corpuses specific to individual scholars rather than few huge collections created by large communities of scientists (e.g., Genome databases, Earth Observation System databases). Thus, in the humanities, it is especially critical that specialized and multiple indexes be created and maintained.

Second, information resources in the humanities are less sensitive to time than resources in the sciences; although some searching in the humanities may be limited by period, the temporal range is typically wide. Thus, finding aids and interfaces may not be able to easily leverage time constraints. Additionally, these indexes and guides themselves must evolve as word usage evolves over time.

Third, humanities resources are often multi-lingual. Individual works may use expressions from multiple languages and resources related to a topic or artist may be available in multiple languages. Since English is a de facto language for science and technology, most of the

discovery tools are specific to English (although statistical retrieval techniques such as latent semantic indexing and n-gram analyses (e.g., Cohen, 1995) have proven generalizable across multiple languages). Machine translation research that uses an interlingual language (e.g., Dorr, 92/93) may also prove useful for indexing multi-lingual corpuses.

Fourth, data acquisition and digitalization is expensive and time-consuming. Simply adopting a controlled vocabulary such as the AAT is a significant change for cataloging new acquisitions, but the retrospective conversion of local cataloging records is intellectually challenging (and controversial) as well as expensive. Also, digitizing text is challenge enough, but much of the content of the humanities is graphical, aural, and three dimensional. Capturing and storing images or sound at high resolutions is both time consuming and open to criticism vis-a-vis interpretaveness. Furthermore, the compression scheme used will determine/limit what surrogates can be made available for browsing.

Research Needs and Directions

The special challenges the humanities offer for search and discovery research and the continued evolution of the Internet suggest several themes for future research and development.

Multiple Approaches. Because humanities scholars typically do not look for answers to well-defined questions but rather elaborate threads of discourse, traditional database techniques will not suffice. There is a need for humanities scholars and communities to create and share thematic indexes specific to their own interests and expertise. The metaphor of self-organizing systems--many minds creating entry points for search and discovery--seems more appropriate both for a world-wide network of information and for the spirit of the humanities than the top-down metaphor of one great mind/committee that provides an organizational framework for some master index. Because it is in their personal interest to create such thematic indexes, humanities scholars will do so without funding (funding will speed up this process). There are, however, two crucial needs for research support in this regard.

First, we must learn how to aggregate thematic indexes and forge links among them that activate according to the ontological perspective of the information seeker (this may be thought of as a kind of intellectual interoperability). Thus, information seekers can specify a Òschool of thoughtÓ and be given sets of links that are customized to that perspective. Another user with a different perspective would find a different set of links for the same corpus. Research in thesaurus merging (Rada, 1985), scheme merging (Nica & Rundensteiner, 1995), and ontology definition (Wiederhold, Wegner, & Ceri, 1992) may eventually be helpful here.

Second, scholars should be encouraged to create pathfinders--guides to themes or topics that not only give pointers to information resources but also critical commentary and interpretations about those resources. Since it is likely that we see the continued development of independent, non-standard collections of information--each celebrating human innovation and creativity through unique organization and expression--it makes sense that these collections themselves should become subject to study, critique, and interpretation. Thus, the purposeful aggregation and added-value commentary that define pathfinders in the humanities represent a form of scholarship that deserves directed research attention. Commentaries have long been part of

scholarly practice in the humanities but electronic environments provide new possibilities for creating critical threads through the electronic morass that themselves may include interactive aspects; e.g., using a pathfinder a second time will be different since it will take advantage of knowledge about what you have already examined. How this knowledge is used requires creative and scholarly decisions on the part of the creator of the pathfinder.

Because Internet resources will be available to a broad range of users from children to seasoned scholars, there must be simple as well as powerful tools for search and discovery. Although these are not mutually exclusive requisites, there is a need for developments of progressively powerful tools as well as tools tuned to specific types of users (see also Murray). A related need is for systems that provide multi-lingual interfaces as well as search and discovery tools that handle multi-lingual corpuses. Both of these needs have positive implications for humanities since they will lead to new classes and groups of users.

Other Needs. Clearly, there is a need for more materials in the humanities to be transferred to electronic form (see Kenney). Especially for text-based fields, techniques for automatically categorizing and summarizing text fragments will be necessary if information seekers are to maximize their time and memory resources when examining and scanning candidate texts. It seems prudent to look for ways to combine statistical approaches with knowledge-based approaches. For image-based fields, techniques to extract and match patterns must be combined with whatever word-based information is available (see Romer). Regardless of the medium (text, audio, images), interface mechanisms that allow rapid scanning (e.g., zooming and panning; fast-forward, multiple display panels, etc.) are essential to an integrated search and discovery environment.

Finally, scholars must consider their audiences both during and after creation of work. First, during creation, the work can be tailored to make it more easily found by the audience. On the crass side, this is advertising before art; on the scholarly side, this is tailoring expression to be best understood by oneÕs public. Second, after creation, the scholar can point the work at audiences. This is what publishers presently do but a networked world allows creators to broadcast or narrowcast as they please. This closer link between creators and consumers depends on development of tools that support creation, communication, and maintenance of digital work. (We can imagine next iterations of hypertext authoring systems such as Storyspace that automatically generate html and browser scripts that monitor usage statistics for automatic (or random) mutations or author version control.) Surely, tools will emerge that allow creators to produce ÒviralÓ works that change depending on use (or alternatively, appear in different forms in different environments). Persistence and stability enables static indexing and locational aids to work in todayÕs libraries. We need research to determine how to document, find and use new genres of interactive and evolving intellectual products.

References

Ackerman, M. (1994). *Answer Garden: A tool for growing organizational memory*. Unpublished doctoral dissertation, MIT.

Coalition for Networked Information (in preparation). *Networked information discovery and*

retrieval (written by Lynch, Summerhill, & Preston). ftp.cni.org.

Cohen, J.D. (1995), Highlights: Language- and domain-independent automatic indexing terms for abstracting. *Journal of the American Society for Information Science*, 46(3), 162-174.

Crane, G. (1992). *Perseus 1.0: Interactive sources and studies on ancient Greece*. New Haven, CT: Yale U. Press.

Dorr, B.J. (1992/3). The use of lexical semantics in interlingual machine translation. *Machine Translation*, 7(3), 135-193.

Getty Art History Information Program, American Council of Learned Societies, & Coalition for Networked Information. (1994). *Humanities and Arts on the Information Highways: A Profile*. Santa Monica, CA: Getty Art History Information Program.

Jones, S., Gatford, M., Rugg, G., Hancock-Beaulieu, M., Robertson, S., Secker, J., & Walker, S. (1995). Interactive thesaurus navigation: Intelligence rules OK?? *Journal of the American Society for Information Science*, 46(1), 52-59.

Lavin, M. (1992). Researching visual images with computer graphics. *Computers and the History of Art*, 2(2), 1-5.

Marchionini, G. (1995). *Information seeking in electronic environments*. NY: Cambridge U. Press.

Nica, A. & Rundensteiner, E. (1995). Uniform structured document handling using a constraint-based object approach. In *Proceedings of Advances in Digital Libraries 1995*. NY: Springer-Verlag, pp. 42-59 (preliminary version).

Voorhees, E. (1994). On expanding query vectors with lexically related words. *Proceedings of the Second Text Retrieval Conference (TREC-2)*, pp. 223-231. NIST.

Wiederhold, G., Wegner, P., & Ceri, S. (1992). Toward megaprogramming. *Communications of the ACM*, 35(11), 89---99.

Thanks to David Bearman, Gregory Crane, Elli Mylonas, and Michael Neuman for comments on a previous version of this essay.