a key area where AI can contribute to digital library development. We have found structured description languages useful for representing agent capabilities—because they directly support construction of descriptions from parts of others—and for reasoning about the relationship between descriptions based on the structure of their parts. Using these facilities, for example, agents can negotiate about provision of complex services, as long as they have a common understanding of the primitives making up those services. For example, in our current prototype, the query planner uses knowledge about the relationship between the index terms of the registry and the taxonomy server to enlist both of these agents for an iterative search for collections "near" the user's query.

We have also found agent communication languages based on speech act categories (as in KQML) quite useful for describing protocols and agent communication strategies. In particular, we have identified a number of speech acts (and appropriate semantics) that cover a broad range of information services. In addition, we are introducing speech acts to cover negotiation: the process by which a set of agents comes to terms on provision of information services and allocation of resources to the various service activities. Our protocols support a general negotiation model, based on agents tendering offers to buy or sell services, basic resources, and other information goods for specified prices. Market facilitator agents help resolve these offers into deals among agents. The specific protocols employed depend on the agent interaction. For example, where the information services are easily parametrized and interchangeable, a competitive auction process involving many agents will be effective. For more individualized services, the negotiation facilitator might need to account for strategic reasoning on the part of participating agents.

As all the hype surrounding software agents attests, digital libraries are just one arena for a new generation of automated information services. AI has a large role to play in improving the generality, robustness, and overall competence of these services. We believe that AI has an equally important role to play in architectural infrastructure (in concert with other technologies, of course). In a digital library, our typical problem is an abundance of available information and information services. Efficiently bringing together the right agents with the right resources for the right tasks is the measure of the library's effectiveness.

References

1. D.E. Atkins et al., "Toward Inquiry-Based Education Through Interacting Software Agents," *Computer*, Vol. 29, No. 5, May 1996, pp. 69–77.

Digital librarians: beyond the digital book stack

Jaime Carbonell,
Carnegie Mellon University

What exactly is the purpose of a traditional library in our society? What useful services does it render? What additional services should it render if only such services could be made cost- and time-effective? Without facing these questions squarely, the endeavor of creating digital equivalents of our present libraries may not yield the most productive results. At the risk of oversimplification, the primary *primary* functions of a library, as perceived by an avid user, are

- Archiving large collections of primarily textual materials (books, journals, and encyclopedias—and, increasingly, sound and video recordings, as well).
- Indexing archived materials by subject, title, author, date, and so forth, to permit unaided user search and access.
- Providing reference librarians to aid in seeking information when unaided search proves inadequate and to perform a number of related and more sophisticated functions.
- Offering acquisition librarians to determine what new materials to incorporate and what existing materials to discontinue.

Current efforts at digitizing the library, Carnegie Mellon's Informedia project included, focus almost exclusively on the first two. That is, they concentrate on the library functions and not on the librarian functions. The task of massive digitization, storage, indexing, search, access, and otherwise managing the archive and its access is truly daunting, and well worth the effort. Digitized libraries can be much larger than physical ones (cheaper storage and faster access), more permanent (CD-ROM lasts virtually forever), provide wider access (through the Web as opposed to local commuting distance), and possibly avoid the resource deadlock problem (no more infuriating "checked out" or "lost item" status for just the material you are seeking). But, despite these advantages, let us not forget our helpful librarians. They, too, can add value and make digital libraries truly useful and user-friendly.

The digital reference librarian

Mary, a sixth grader from Central High School, has been given her first truly independent research paper assignment. Being rather courageous, she has chosen to investigate the impact of drug abuse on inner-city school children. Since, like most educated children in the year 2001, she is well versed on accessing the Web, she points her superscan browser at http://www.digitalib.edu and types "drugs" when prompted for a search topic.

Assuming her parents are not the censorious type—that they have not blocked everything with the keywords "sex," "drugs," and "violence," and all relevant synonyms thereof—she gets a zillion hits, properly ranked. The first one is a lengthy US Federal Drug Administration report on generic over-the-counter drugs. So is the second one. And the third one. (They mention "drug" and "drugs" and synonyms such as "pharmaceuticals," and statistically-correlated terms and phrases such as "testing," "over the counter," and "prescription" with very high frequency, much to the

Interesting URLs

Stanford University Digital Libraries Project

http://www-diglib.stanford.edu

Carnegie Mellon Informedia project

http://www.informedia.cs.cmu.edu

University of Michigan Digital Library project

http://www.si.umich.edu/UMDL/

***Computer*'s May 1996 special digital libraries issue**

http://www.computer.org/pubs/computer/computer.htm

Coming in August

Neural Networks in Real-World Applications

"Neural Network Speech Processing for Toys and Consumer Electronics"
Michael C. Mozer, Sensory Circuits, Inc. and University of Colorado, Boulder

"How to Make Use of Neuroinformatic Processing in Real-World Face-Recognition Applications"
Wolfgang Konen , ZN GmbH, Bochum

"Neural Networks Provide Robust Character Recognition for Newton PDAs"
Larry Yaeger, Apple Computer

"Neural Networks for Computer Virus Recognition"
Gerald J. Tesauro, Jeffrey O. Kephart, and Gregory B. Sorkin, IBM Thomas J. Watson Research Center

liking of the sophisticated search engine.) Mary looks at these reports with alarm, and skips down to the fiftieth one in the list. That one, finally, is not an FDA report; it deals with a comparison study of analgesic drugs for treating arthritis.

Frustrated, Mary goes back to refine her query—she is really good at Web surfing—and types "drugs that are very bad for you" and gets pretty much the same result (the other words being mostly "stopwords"—very high-frequency nonspecific words set aside by the sophisticated 2001 search engine). She tries once more: "kids that get all messed up because their parents let them smoke or pop what they like, or because they do it anyway and their parents don't know and don't stop them, and then they get in trouble with the school and bad grades and all that and maybe even drop out."

Aha! the super-duper search engine really goes to work on this one, focusing on good terms like "kids" and "parents" and "school" and "grades" and "not know" and "drop out"—clearly a query about the education system and its institutional failings. Mary gives up, and is ready to tell the teacher there is nothing on drugs in schools in the much-vaunted digital library.

Should we blame Mary for not formulating a better query? The search engine for not divining Mary's intentions? Or, the lack of a reference librarian? I contend the last. Had Mary suffered a similar frustration at our public library, she would have received help in the form of a clarification dialog with the reference librarian and would have been told to ask for "illegal drugs in inner-city elementary schools" or would even have been provided with direct pointers to the books and articles she sought—at a fifth-to-eighth-grade reading level to boot.

Let us then not just build digital libraries but also digital librarians. A digital reference librarian (DRL) starts with an interactive dialog interface to elicit the information needs of users—especially users unsophisticated in library search. In the example just cited, the DRL would first read any user profile information (11-year old, frequent Web user) and focus on refining the query—"What type of drugs: over-the-counter (aspirin or Tylenol), prescription (such as penicillin), illegal drugs (marijuana or cocaine)?" And, on selecting the last, it would ask the user to say more about the topic, then proceed to refine again, for instance, to "the consequences of drug use in elementary schools."

Such DRLs can and should be built, patterned after experienced human librarians. Of course, not all functions of human reference librarians can be easily automated, but interactive elicitation of information to formulate meaningful queries is a very useful start. Moreover, such a task does not require in-depth knowledge of the subject domain, but rather knowledge of the search engine (what kinds of words, phrases, and concepts are useful), the categories of information organized in an ontology (such as multiple types of drugs), much like the present cataloging system; other selection criteria (for instance, reading level, recency of information, or diversity of sources); and the ways to ask questions whose answers the DRL can directly interpret.

The digital acquisition librarian

How do we allocate our library acquisition budget? Do we subscribe to new journals and drop the ones with the least circulation? Which new books do we buy and catalog? These are central questions to librarians, best answered if accurate information-need profiles are available. Of course, in traditional libraries, such profiles are never available in statistically meaningful ways. At best, circulation profiles of existing materials can be accessed and used to extrapolate future needs based on presently serviced information needs. But, there is no hint at what kinds of information needs go unmet, other than the occasional anecdote.

As we build digital libraries and DRLs, we can do better. All user queries and system responses become available as data (dropping user names if anonymity is desired). This data contains not only circulation profiles, but topics queried, successful searches, unsuccessful ones, and so forth. In particular, we can find which queries failed to yield desired pointers to the digital library collection. We can analyze this information with existing inductive and statistical data-mining techniques such as nearest-neighbor searches, clustering, frequency-weighted correlation analyses, and time-series trend detectors. Analysis results will point to the areas of unmet user needs, because either the materials are not in the digital library (giving strong hints to the acquisition librarian), or they are present and either the indexing structure or search engine failed to make the connection (pointing to needed improvements).

Beyond traditional library functions

Of course, as we envision future digital libraries, we need not be constrained to those functions and services provided by present-day traditional libraries. The discussion thus far focused on not giving up useful functions, such as reference librarians, as libraries go digital, and on exploiting the digital medium to improve these functions, as in automated analysis of unmet user needs to drive acquisitions.

Digital libraries can transcend traditional ones. Some ground-breaking approaches are already part of the digital library effort, such as multimedia access and news-on-demand at the CMU Informedia project. The paragraphs below sample the most interesting potential developments, in my opinion, starting from more modest technological ideas to large-scale paradigm shifts.

Subdocument search and indexing. Libraries index entire documents, not sections or passages within. However, in information retrieval there is significant work on finer-grain indexing, which should be a natural part of a digital library to provide higher-fidelity information access.

More sophisticated information retrieval. Rather than focus simply on the tried and true metrics—precision (what fraction of the retrieved documents are relevant) and

recall (what fraction of the relevant documents are retrieved)—we need new measures as the scale of digital libraries increases. For instance, we need to measure novelty of information: retrieving 100 relevant documents that say pretty much the same thing is no better than retrieving one relevant document. We also need to measure appropriateness of information to the user: juvenile readers might not comprehend detailed technical treatises. We further need to weigh whether the user wants everything directly pertinent to a query—as in a lawyer requiring every potentially on-point precedent case to defend against a massive lawsuit—or whether sampling and diversity are greater virtues, given the user's information needs and profile.

On-demand document summarization. Articles might have abstracts and books might have summaries. But, these are created by the writer without knowledge of the specific information need of a user searching the library. Researchers at CMU are developing query-relevant summarization technology that synthesizes variable-grain summaries of the information in a document that pertains directly to the query.

Multidocument summarization. Ideally, users should receive a summary of everything pertinent to a (narrowly defined) topic or query, transcending document boundaries, without redundancy of information contained in multiple documents. Moreover, users should be able to zoom in and out from headline summaries to page-long ones to report-length topic summaries. If the summary provides insufficient detail, the user should be able use a single mouse click to access the original source documents from whence a particular passage in the summary was extracted.

Multimedia search and indexing. A multimedia digital library requires not just standard indexing and retrieval, but also subdocument indexing and summarization technology—more than do paper documents. It is harder to "flip through" a film (other than the confusing fast-forward, which loses the sound track), and films seldom contain tables of contents or indexes.

Access to live or near-live information. Why stop at archival information? Libraries often have a periodical reading room, but this material is not always indexed in the card catalog or its electronic equivalent, because it is too costly to index live information. With automated indexing, there is no need to treat live information—newspapers, newswire feeds, radio and television news, stock quotes, and so forth—as a second-class category.

Active information sources. And why stop at passive information that can be read, heard, or seen? Interactive programs—educational software, investment counselors, interactive how-to manuals, medical advice givers, and so on—can and should be part of the digital libraries, fully exploiting the interactive computing substrate.

Symbiotic human-machine gurus. And why stop at interactive programs? Why not include on-demand human advice (such as when the medical advice program is stumped), if available? The possibility of human expertise fully integrated into a passive and active information medium could vastly expand the concept of a library. (Of course, issues such as payment for service and certification of expertise must be resolved.)

Information on tap, anywhere, anytime. Fortunately, the physical-access side of the equation is already under intensive development. Thanks to the Internet and Web technologies, and forthcoming improvements in wireless networks, pen-based interfaces, speech recognition, and related areas, ubiquitous access to the digital library will not be a problem. (Now, how do we use the digital library to entice Johnny to read more?)

The universal library. All these developments indicate that the concept of a digital library is really a transitory phase toward the *universal library*—a vast distributed information and active-advice repository accessible from anywhere with increasingly improved indexing, extraction, and summarization techniques. It will be a library without walls or national boundaries.

A digital librarian is essentially a special type of automated agent that combines functionality for information elicitation, planning, data mining, and coordination of search, retrieval, and content summarization.

- Elicitation requires further research in task-directed knowledge acquisition—for instance, how to ask questions whose answers are directly interpretable for the task of query reformulation.
- Search and retrieval coordination requires explicit modeling of the retrieval engine and indexing structure so that the agent has a goal of creating informative queries with respect to the search system.
- Content summarization facilitates user feedback on whether the correct resources have been located. It requires research on dynamic extraction of key information pertinent to the query or topic, rather than simply providing a context-free summary much less relevant to the task.
- Finally, digital librarians will require a limited form of data mining to extract unmet information-need patterns from accumulated user queries. For this purpose, unsupervised learning techniques, such as clustering and composite-term discovery techniques, are directly relevant, rather than the more traditional inductive supervised learning for automated data classification. We do not know ahead of time which topics or areas not covered by the digital library might be of interest—that's precisely what we are trying to discover.

Advances in all of these technologies are underway, but not yet coordinated and targeted at the task of creating a digital librarian.