Searching has come a long
way since the 60s, but
have we only just begun?

From IR to Search **AND BEYOND**

BY RAMANA RAO, INXIGHT SOFTWARE

It's been nearly 60 years since Vannevar Bush's seminal *Atlantic Monthly* article, "As We May Think," portrayed the image of a scholar aided by a machine, "a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility." Unmistakably in this is the technology now known as *search* by millions and known as *information retrieval* (IR) by tens of thousands. From that point in 1945 to now, when some 25 million Web searches an hour are served, a lot has happened.

In the mid-1980s at Xerox PARC I witnessed the beginnings of a research effort related to search that has swept me along for nearly 20 years. By that time, search and the desktop metaphor had become serious commercial forces. It was also clear, at least to many researchers, that both search and graphical user interfaces would reach their limits as the amount of networked information grew and as a broader range of users and uses became common. Yet, rapidly increasing processing power and graphical capabilities would allow us to build information workspaces that could go much further in allowing people to use personal, organizational, commercial, and public information.

In this article I take you on a tour of search history, beginning in the 1960s, to that point at PARC in the 1980s, and then to mainstream uses of information currently on the Internet. Across the decades, two dichotomies stand out clearly:

From IR to Search AND BEYOND

- The first is the contrast between focusing on the one hand on narrowly defined technological approaches and on the other hand on a broader understanding of the full problem set and the possible solutions.
- The second is the contrast between working out ideas in research versus spreading them commercially.

As someone who has stood in one camp or the other at various times, I don't see these as either/or choices, but rather as orientations that each support the longer-term goal of making things better for lots of people. Progress requires perfecting approaches and technologies for solving constituent problems, as well as engineering technologies that fit properly into real work and that can be adopted in the real world.

1960S: THE MODEL-T OF INFORMATION RETRIEVAL

The post-war period of the 1950s was marked by furious progress in science and computing, and along with this a dramatic growth of scientific literature. Faced with the challenge of organizing the growing base of scientific content with its rapidly expanding vocabularies, librarians and information scientists grappled with applying

cataloging and indexing theories. Meanwhile, information and computer scientists started to explore mechanized support for both indexing and retrieving content.

Information retrieval, coined as a term in 1952, started picking up speed as a discipline in 1958 at the International Conference on Scientific Information. Not surprisingly, given the context of library metaphors, the basic architecture for IR systems is based on two primary functions that correspond to the traditional activities of organizing a library and finding documents in the library. Also not surprising, the model of how users interact with a retrieval system resembles the traditional model of interaction with a librarian. Users say what they want and the system delivers it.

Framing this more precisely to support system building, the model is that the user—with some information need—fashions a request as a query, and the system returns documents with relevant content as a results list. In short, the model (see figure 1) is QIRO (query in, results out).

From the beginning, leading researchers well understood the challenge of reducing this model to practice. Both the user and the system must operate in ways that only approximate the ideal model. Users usually don't fully understand their own information needs in advance, or else they can't express their needs in a manner suitable for the system to process. The system, thus lacking a complete query or any real understanding of documents, can't effectively match against relevant documents. The surface variability and ambiguity of human language only increase these difficulties.

This inherent challenge of how to match query to output led IR researchers to focus on relevance ranking in which results are ordered according to a degree of matching. While Boolean matching is conceptually straightforward with structured tables of relational data, it's a completely different matter for documents expressed in richly structured natural language.

During the 1960s the framework for evaluating IR systems was also settled. It was based on two key metrics: precision and recall. *Precision* is the percentage of documents in your total returned results set that's actually relevant to you; your return set may contain 100 documents, but the system has low precision if only 15 of those returned documents are relevant. *Recall* is the percentage of all relevant documents that are actually returned; your return set is 12 documents, but you know that another 5,236 relevant documents are out there. Intuitively, precision is about how clean the results set is, whereas recall is about how complete the results set is. These two measures

tend to be inversely correlated, and a system could be biased toward one or the other (see sidebar on this page).

Search systems in the form of online catalogs were first commercially available in the 1970s. These early online search systems—Dialog, for example—focused on searching bibliographic records, references, or surrogates rather than actual documents. They used Boolean query languages, which increased the burden on the user of the system, typically a librarian. Full text systems became available only late in the 1980s—and relevance ranking has roared ahead in the 1990s with Web searching. Through all this, the fundamental QIRO interaction model remained largely intact.

Even in the 1960s, however, a number of approaches related to broader tasks or styles of interactions—including categorization, summarization, extraction, and visualization—were suggested and actually pursued in research. Bush's article pointed out that information wasn't found in libraries because of "the artificiality of systems of indexing" and suggested "associative threads" as a more powerful way to interact with content.

1980S: USER POWER UNLEASHED, A NEW MODEL EMERGES

In the 1980s, personal computing really took off, and following closely after came networked computing and the graphical user interface, along with the desktop metaphor. This metaphor focused largely on applications as editors for programs, documents, drawings, and so on. It also focused on supporting networked communication and access to file, print, and directory services. With networked personal computers widely deployed throughout Xerox, it was quite easy to see the coming challenge. As large collections of documents appeared and the network grew, not just inside Xerox but also outside on the Internet, finding files or resources was becoming difficult. Often, you had that haunting feeling that somewhere out there were people or documents that could save you a great deal of work.

It was easy to foresee a shift from document creation to information access—and to realize that the navigational interface of the desktop metaphor wouldn't work well for finding relevant documents as they proliferated across networks. Looking at the information retrieval research and systems available at that time, it was equally clear that the QIRO model had its limits. Besides the inherent challenge already outlined, other difficulties would arise as a broader range of users and applications needed to be supported. In particular, the QIRO model ignores a number of realities of information work, espe-

Classic Concepts of Search

Architecture
Indexing/searching
Interaction
Query in, results out (QIRO)
Result matching
"Relevance"
Ranking evaluation metrics
Precision/recall

cially in the context of networks and personal computers:

- Retrieval is naturally interactive, iterative, and interleaved with other activities. Often the process of searching sharpens users' understanding of their information needs and the best ways or places to search.
- Users aren't trying to find documents per se, but rather to use the documents to fulfill some broader task. Retrieval is embedded in processes of understanding and analyzing information that are in turn embedded in still broader processes of creation, learning, planning, operating, and decision making.
- Users need to access many disparate collections, with varying characteristics of provenance, authority, quality, coverage, and form. Most personal and organizational collections are naturally messy accumulations of highly varying documents, and there is little time or resource to organize or curate them.
- Search services and software vary widely in functionality, performance, interfaces, economics, and availability. Effective retrieval depends on users forming effective search strategies over the space of possibilities, considering characteristics of the source (collections and service) and contextual factors related to task and setting.

The Intelligent Information Access project at PARC formed a vision of fusing the QIRO information retrieval model and the graphical desktop metaphor into an information workspace, as illustrated in figure 2. Increasingly, computation could be used for creating richer illusions, doing more sophisticated content analysis, and supporting richer dialogues by tying these processes together.

In the PARC model, the user, engaged in larger work processes, manipulates objects in the workspace, retrieving units of information from multiple disparate sources. This model focused on a number of key ideas:

- **Search and browse.** The information workspace would support not just search but also browsing. These two styles of dialogues have complementary strengths and weaknesses. Each can be used in different kinds and

From IR to Search AND BEYOND

stages of tasks. For example, QIRO style dialogues can be quite efficient and effective when they work at all, while browsing can be easier to learn and use in many cases.

- **Docuspace and concepts.** The information universe includes the whole hierarchy, from all sources to whole collections, document lists, and documents, including document sections, sentences, and unit concepts. Other important distinctions include the dimension of personal, organizational, commercial, and public information and the dimension of messy accumulations to highly crated collections.
- **Maps and digests.** Visual maps of information spaces enable both the understanding of regular and unique patterns and relationships across the large numbers of objects—whether they represent sources, documents, or facts. In addition, well-composed previews, digests, and summaries of results and documents can guide users to the most relevant items and facilitate a rapid understanding of what's found.
- **Indexing and extraction.** Indexing to support typical search dialogues is a relatively impoverished form of content analysis. Other content analysis techniques, based on linguistic analysis and statistical techniques,

offer great promise for tagging content with meta-information that could be used in organizing collections or browsing dialogues based on maps and digests, as well as in new kinds of text-mining applications.

- **Memory and reuse.** Access is a process that takes place over long periods of time, so both historical capture and reuse of previous strategies can be quite valuable. The best approach is to allow an incremental refinement and reuse of past work as new activities merit the ongoing attention required. Thus, history, process, and search management are important pieces of functionality in an information workspace.

Researchers explored these ideas at PARC and perhaps a dozen other places right through the 1990s. While commercial efforts for the most part were made to provide richer workspaces and better IR functionality, the commercial world focused on simple interactions to enrich new services and information on the network. The QIRO model saw an explosive success with Web searching. Now, more than 10 years later, millions of users are familiar with the limitations of simple search and are seeing the real commercial uptake of the broader information workspace ideas.

1990S: SEARCH MAINSTREAMS, NEW MODEL COMMERCIALIZING

The 1990s, which indeed supported our predictions in the 1980s, saw a proliferation of information sources available from desktop computers. In addition to the documents created and managed by individuals and their workgroups, huge numbers of documents are now available from servers within enterprises and on the Internet. Furthermore, commercial and public online information sources that provide access to bibliographic citations, newspaper and magazine articles, financial and business data, and much more have only expanded further with the Internet.

Whereas the original commercial thrust for search in the 1970s was directed at online services, later efforts offered search as a software package to be applied to personal or organizational content. In the 1990s, with the expansion of the Internet and intranets, both vectors were pushed forward. A new breed of online search service in the form of Web search engines concentrated on search over full text and the truly wide-ranging and messy collection available on the public Web. In parallel, enterprise search, offered in the form of client-server software, became more common to support access to internal Web servers and document repositories.

Neither of these two prongs of search in the 1990s

moved beyond the QIRO model. There certainly was some attention given to improving relevance with full-text searches, but not much attention was given to improving or expanding the QIRO model. Rather, commercial efforts primarily focused on broader deployment and business concerns. In the case of Web search engines, the focus was on coverage, latency, scale, and other issues related to offering search of public Web content. Meanwhile, the primary concerns of enterprise search related to the typical enterprise software concerns of providing complete IT functionality for administration, integration, customization, client-server architecture, APIs, security, and so on.

Interestingly, a number of the Web search businesses—in fact, the successive Web search leaders since 1995 (Infoseek, AltaVista, Inktomi, Fast, and Google)—have all tried to cross the firewall by offering packaged versions of their Web search engines to enterprises. Not surprisingly, when comparing the primary concerns of the online services and the enterprise search products, none of these appear to have wiped out the incumbent enterprise search product leaders.

In the last few years, it's easy to see many of the information workspace ideas being absorbed into commercial efforts. As the technology design ideas are being commercialized, they are being “widgetized” or packaged into market categories of functionality that include the following:

- **Advanced search:** A number of companies (including traditional search companies) have started to incorporate more sophisticated indexing/matching algorithms—many quite old, including automatic query expansions—as well as linguistic and statistical techniques for dealing with language variability and ambiguity (e.g., latent semantic indexing).
- **Categorization:** The first widespread nonsearch functionality is categorization, which supports the automatic populating of searchable and browsable information directories, typically called *taxonomies*, and the creation and management of the taxonomies.
- **Extraction:** Linguistic content analysis can be used to pull particular elements out of documents. Two especially valuable types of extraction are entity extraction and fact extraction. Entity extraction involves pulling out proper noun phrases (e.g., organizations, people, and places). Fact extraction includes identifying relationships among these entities, understanding the roles played by various entities, and identifying key events.
- **Visualization:** Interactive tools beyond conventional user interface widgets provide an overview, as well as

navigation at all levels—from the whole universe to particular collections, to results sets, right down to the elements of documents.

- **Metasearch and federated search:** Search can be supported over multiple collections in a variety of ways, most notably by *metasearch*, providing a search of

Visit these Web sites to find out more about new visions of search.

CLUSTERING RESULTS

WiseNut

<http://www.wisenut.com>

Dogpile

<http://www.dogpile.com/>

Visualizations

KartOO

<http://www.kartoo.com>

TouchGraph

<http://www.touchgraph.com>

Groxis

<http://www.groxis.com>

Inxight

<http://www.inxight.com>

METASEARCH

Dogpile

<http://www.dogpile.com/>

infoseek

<http://infoseek.go.com/>

Intelliseek

<http://www.intelliseek.com>

SEARCH AND CATEGORIZATION

Convera

<http://www.convera.com>

Inxight

<http://www.inxight.com>

Verity

<http://www.verity.com>

From IR to Search AND BEYOND

models from each collection to find appropriate collections, and by federated search, brokering queries to multiple search services and combining the results.

- **Summarization:** Extracting key sentences from documents is considered by many as a way to understand a particular document, but it's increasingly common to find applications that compose views of extracted information over not just one document, but results sets and whole collections.

Both Web search services and enterprise search products are incorporating one or more of these functionalities. Though the larger Web services and software products are typically more conservative, I believe they will either absorb the ideas or be surpassed by those that do. Many of these ideas can be tried on the Web sites listed in the resources section of this article.

2020 FORESIGHT

The last 60 years of search have seen ideas from research eventually get adopted in mainstream commercial settings. Although immediate costs and business requirements may have forced initial commercialization efforts into limited versions of the research ideas, ultimately the exponential growth of computing power and the pressures of supporting a broader audience have driven the adoption of a fuller set of ideas. The following four predictions essentially lay out a richer, broader, more uniform model of information interaction that I believe will become a standard part of educational, cultural, and organizational realities over the next 15 years.

RICHER USER MODEL OF INFORMATION SPACE

A large mainstream audience will share a rich conceptual model of the information universe. This model is already common among many who actively use networked information. A central aspect of this model is the essential hierarchical organization of information into universe, libraries, collections, documents, document parts, sentences, concepts, and objects. Crossing-cutting this essentially hierarchical layering is a variety of relationships that will be commonly understood, including refer-

ences, attribution, and versioning. One key aspect is the understanding of the role of meta-information at each level, which is as important to the use of the information as the information content itself.

This model will undergird a common standard of information literacy, and a new set of skills will be necessary to survive and thrive in the new networked information urbanity of the future. Questions such as "Where in the universe should I be searching?" and "How should I navigate through the universe accumulating the information I need?" will be answerable in this broader conceptual framework.

RICHER FUNCTIONS FOR INFORMATION USE

Just as the QIRO model has become mainstream with the spread of Internet technologies, so too will the information workspace model. Interaction in the information workspace will be based on three new constructs.

Text mining will catch and eventually dwarf traditional information retrieval.

- **Maps.** As is the case with physical maps, conceptual and perceptual maps of the universe, collections, and documents will become resources for both understanding overall structures and navigating to specific areas of interest.
- **Digests.** Well-designed digests will provide "a little bit, but not too much" information about any objects at all levels of the information hierarchy.
- **Extractors.** Operators for analyzing content will allow users to explore text and discover relationships and patterns, as well as unusual or unique occurrences.

Retrieval systems for public, commercial, and private content will all adopt standard maps, digests, and extractors. Essentially, as our shared ontologies of information space become more sophisticated, so too will our expectation of information access functionality.

RICH INFORMATION WORKSPACES BASED ON OPEN INFRASTRUCTURE

Our information workspaces will finally achieve the richness, flexibility, and naturalness of our physical

workspaces, while integrating digital reach and augmentation. These workspaces will support both individual and collaborative information activities, smoothly integrating information access with information processing, synthesis, and analysis. The workspace will be open, allowing for the easy assembly of standard, common, specialized, and customized elements—maps, digests, and extractors—and will have access to wide varieties of sources along with standard models of those sources.

The eventuality of openness is supported by the broader picture of IT evolution as driven by the growing cost and increasing competitive pressures on large organizations. An open workspace will be possible because of standardization around software environments that allow a flexible integration of interface, communication, computation, and content components and services. Open source and the emerging hosted models will play out for information access functionality as they are for other areas of software functionality. All of these factors, along with the limits on the complexity of large-scale and broad-audience solutions, will drive consolidation toward a standard set of services and standard widgets, view types, and dialogues for information access.

GRANULAR USE OF LINGUISTIC STATEMENTS

For 60 years search has focused on helping users retrieve documents. This use of computation is a “pave the cow path” model based on the traditional physical containers for information and the traditional model of retrieval in libraries. In such models, the human is left to the task of scanning, reading, digesting, and otherwise assimilating the contents of the book or journal or article. The broader models certainly help form better access strategies, as well as better target documents or document sections that deserve further attention, but there are yet greater opportunities.

Text mining will catch and eventually dwarf traditional information retrieval. This pursuit model starts with the linguistic and statistical text processing being used today in highly valued targeted applications—for example, counterterrorism or drug discovery—without having to overcome the full challenge of natural language understanding by machines. Though I seriously doubt that a full understanding of the problem will be achieved by 2020, more focused applications of text mining will likely become commonplace in this timeframe.

With the rise of text mining, I foresee an intersection of two long distinct histories of computational use, one supporting organizations and the other supporting individuals. Enterprise data computing—embodied by

mainframes, relational databases, ERP (enterprise resource planning), and other enterprise applications—has been the main driver of big IT technology, while personal computing—embodied by desktop environments and applications, communications technologies, entertainment, and other consumer technologies—has supported the individual and collaborative work of humans. I believe that by 2020 the processing of language-based information will surpass the processing of operational data originally captured in structured databases.

The semantic Web may come to pass, but not through the process of humans learning to act as machines, or computers replicating human skills, but rather through the design of whole systems, as suggested by J.C.R. Licklider in 1960, that support human-computer symbiosis. Q

RECOMMENDED READINGS

1. Bush, V. As We May Think.” *Atlantic Monthly* (July 1945); <http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm>.
2. Licklider, J. C. R. Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1 (March 1960), 4–11.
3. Rao, R., Pederson, J. O., Hearst, M. A., Mackinlay, J. D., Card, S. K., Masinter, L., Halvorsen, P.-K., and Roberston, G. G. Rich interaction in digital libraries, *Communications of the ACM* 38(4) (April 1995), 22–39.
4. SearchEngineWatch: see <http://searchenginewatch.com/>.
5. Sparck-Jones, K., and Willett, P. *Readings in Information Retrieval*. Morgan Kaufmann Publishers, San Francisco: CA, 1997. Includes a helpful history by D. R. Swanson called “Historical Note: Information Retrieval and the Future of an Illusion.”

LOVE IT, HATE IT? LET US KNOW

feedback@acmqueue.com or www.acmqueue.com/forums

RAMANA RAO is chief technology officer and a founder of Inxight Software. Throughout his career, Ramana has pursued the design of software that extends the intellectual and creative reach of knowledge workers, mainly because he’s always wanted to be smarter and more creative. At Xerox Palo Alto Research Center (PARC) for 10 years, Ramana did all kinds of great research on intelligent information access, digital libraries, information visualization, and user interfaces. And he writes regularly about how these ideas will matter soon at www.ramanarao.com/informationflow/.

© 2004 ACM 1542-7730/04/0500 \$5.00