Author's Circulation Version, Originally published in ACM Queue, May 2004.

From IR to Search and Beyond

Ramana Rao

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 by tens of thousands. From that point in 1945 to now, when some ** 25 million Web searches an hour alone are served, a lot has happened.

In the mid-eighties, at Xerox PARC, I witnessed the beginnings of a research effort related to Search that has swept me along for nearly twenty years. Already at that time, Search and the desktop metaphor were becoming serious commercial forces. It was clear, at least to many researchers, that both of these 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, from early days in the sixties, to that point at PARC in the eighties, to mainstream uses of information on the Internet now. Across this period, two dichotomies of how best to apply one's effort stand out clearly for me.

- The first is the contrast between focusing on narrowly-defined technological approaches or instead on broader understanding of problems and full solutions.
- The other is the contrast between working out ideas in research versus spreading them commercially.

As some one that has stood balanced on one foot or the other at various times of each of these dichotomies, I don't see these as either/or choices, but rather as orientations that each support the longer 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 can be adopted in the real world.

1960s: The Model-T of Information Retrieval

The post-war period of the fifties was marked by a 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 (Sparck Jones). Not surprisingly, given the context of library metaphors, the basic architecture for IR systems are based on two primary functions, corresponding to traditional activities of organizing a library and finding documents in the library. Also not surprising the model of how a user would interact with a retrieval system resembled the traditional model of interaction with a librarian. A user would say what he wanted and the system would deliver it.

Framing this more precisely to support system building, the model is that the user, with some information need, would fashion a request as a query, and the system would return documents with relevant content as a result list. In short, the model, as shown in Figure 1, is Query In, Results Out (QIRO).

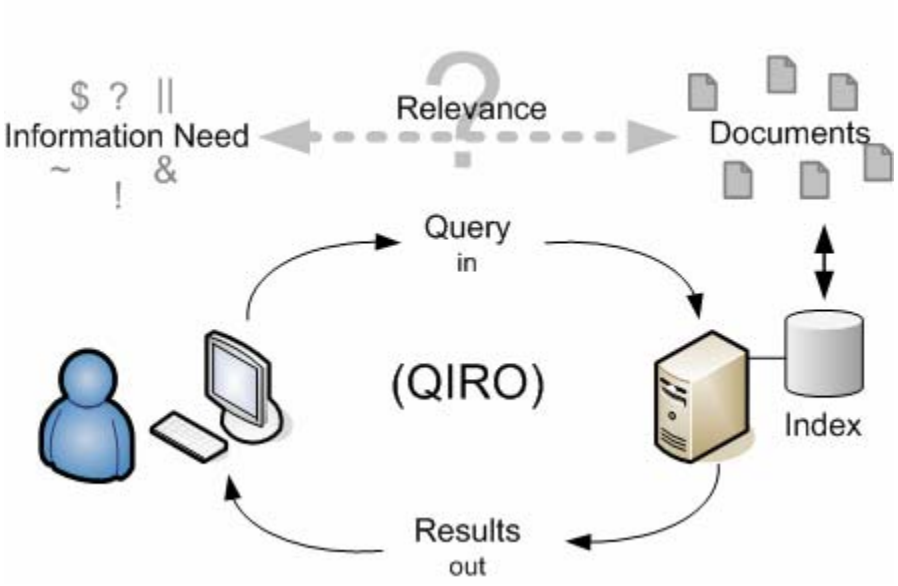

Figure 1: The Classic Search Model of Query In, Results Out

Certainly, from the beginning, the leading researchers understood well the challenge of reducing this model to practice. Both the user and the system must operate in ways that only approximates the ideal model. Usually, the user doesn't fully understand their own information need in advance, or else, he can't express the need in a manner suitable for the system to process. Meanwhile, the system, lacking a complete query or any real understanding of documents, can't match effectively against relevant documents. The surface variability and ambiguity of human language only increases the difficulties.

This inherent challenge, of how to match query to output, lead IR research 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.

Also settled in the sixties was the framework for evaluating IR systems based on two key metrics, precision and recall. *Precision* is the percentage of documents in your total returned result set that are actually relevant to you—your return set may have contained 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—if your return set is 12 documents, but you know there are 5236 relevant documents out there. Intuitively, precision is about how clean the result set is while recall is about how complete it is. It was readily seen that these two measures tended to be inversely related, and that a system could be biased toward one or the other.

The fundamental concepts of classic search can thus be summarized:

Architecture:	Indexing/searching
Interaction:	Query In, Results Out (QIRO)
Matching:	"Relevance"
Evaluation:	Precision/Recall

Search systems in the form of online catalogs first became available commercially in the 1970s. These early online search systems, e.g. Dialog, focused on searching bibliographic records, references or surrogates rather than actual documents, and used Boolean query languages, pushing more burden onto the user of the system, typically a librarian. Full text systems started becoming available only late in the 80s, and relevance ranking has roared ahead in the 90s with Web searching. Through all this, the fundamental QIRO interaction model stayed largely intact.

However, even in the sixties, a number of approaches related to broader tasks or styles of interactions---including categorization, summarization, extraction, and visualization---were suggested and even pursued in research. In fact, even in Bush's 1945 article, he points out that information wasn't found in libraries because of "the artificiality of systems of indexing," and offered "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 it closely, came networked computing and the graphical user interface with its desktop metaphor. The desktop metaphor was focused largely on editors, say for programs, documents, drawings and so on. It was 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 there was a haunting feeling that somewhere out there were people or documents that could save you a great deal of work.

It was easy to foresee that there would be a shift from document creation to information access. Also 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 of the time, it was also equally clear that the QIRO model had its limits. Besides the inherent challenge outlined above, other difficulties arise with supporting a broader range of users and applications. In particular, the QIRO model ignores a number of realities of information work, especially in the context of networks and personal computers.

- Retrieval is naturally interactive, iterative, and interleaved with other activities. Often the process of searching sharpens a users understanding of their information need and the best way or places to search.
- Users aren't trying to find documents per se, but rather to use the documents in order 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, often distal, 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 very little time or resource to organize or curate them.
- Search services and software varies 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 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 to create richer illusions, to do more sophisticated content analysis, and to support richer dialogues through tying these processes together.

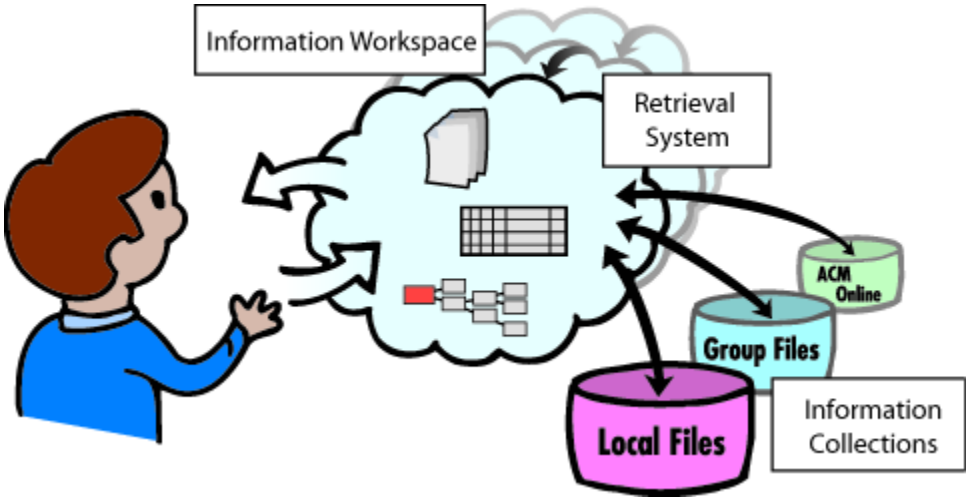

Figure 2: The Information Workspace with QIRO and Many Collections

In the 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 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 easier to use in many cases.
- **Docuspace and Concepts**. The information universe includes the whole hierarchy from the universe of all sources, down through whole collections, document lists and documents, and down into 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 understanding regular and unique patterns and relationships across large numbers of objects, whether they represent sources, documents, or facts. And well-composed previews, digests, and summaries of results and documents can guide users to the most relevant items and facilitate rapid understanding of what is 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, in browsing dialogues based on maps and digests, and 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 all important pieces of functionality in an information workspace.

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

2000s: Search Mainstreams, New Model Commercializing

The nineties, indeed supporting our predictions in the eighties, saw a proliferation of information sources available to all users 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 on enterprise and Internet. Furthermore, commercial and public online information source 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 70s was focused on online services, later efforts offered search as a software package to be applied to personal or organizational content. In the 90s, with the expansion of the Internet and intranets, both vectors were pushed forward. A new breed of online search service in the form of the Web Search Engines focused 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, started becoming more common to support access to internal Web servers and document repositories.

Neither of these two prongs of search in the nineties moved beyond the focus on the QIRO model. Certainly there was some attention on improving relevance with full text search, but not much on improving or expanding the QIRO model. Rather, the real commercial attention went to broader deployment and business concerns. In the case of Web Search engine, 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 typical enterprise software concerns related to 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, Google) have each tried to cross the firewall by offering a packaged version of their Web Search engine to Enterprises. Not surprisingly, comparing the primary concerns of the online services and the enterprise search products, none of these 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 including 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, and linguistic and statistical techniques for dealing with language variability and ambiguity (e.g. latent semantic indexing).
- Categorization -- The first widespread non-search functionality is categorization, which supports the automatic populating of searchable & browseable 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 particularly valuable types of extraction are entity extraction and fact extraction. Entity extraction is about pulling out proper noun phrases e.g. organizations, people, and places. Fact extraction includes

identifying relationships between these entities, understanding the roles played by various entities, and identifying key events.

- Visualization -- Interactive tools beyond conventional user interface widgets that provide an overview as well as navigation at all levels from the whole universe to particular collections to result sets and right down into the elements of documents.
- Metasearch and Federated Search -- Search can be supported over multiple collections in a variety of way, most notably by Metasearch, providing a search of models of each collection to find appropriate collections, and Federated Search, brokering queries to multiple search services and combining the results.
- Summarization -- Extracting key sentences from documents is commonly seen as a way to understand a particular document, but increasingly, it is becoming common to see applications that composing views of extracted information over not just one document but result 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 more conservative, as they typically are, I believe they will either absorb the ideas or be surpassed by those that do. Many of these ideas can be tried on the following Web sites.

Clustering Results
www.Wisenut.com
www.Dogpile.com

Visualizations
www.Kartoo.com
www.Touchgraph.com
www.Groxis.com
www.Inxight.com

MetaSearch
www.dogpile.com
www.Intelliseek.com

Search+Categorization
www.Convera.com
www.Inxight.com
www.Verity.com

2020 Foresight

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

1. 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 references, attribution, and versioning. One key

aspect is the understanding of the role of meta-information at each level that plays as important a role in 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 like where in universe should I be searching, and how can should I navigate through the universe accumulating the information I need will be answerable in this broader conceptual framework.

2. Richer Functions for Information Use

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

Maps -- Conceptual and perceptual maps of the universe, collections, and documents, will like physical maps, become resources for both understanding overall structures and navigating to specific areas of interest.

Digests -- Well designed digests 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 becomes more sophisticated, so too will our expectation of information access functionality.

3. 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, extractors, and 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 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.

4. Granular Use of Linguistic Statements

Search for sixty years has focused on helping users retrieving 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 and 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 very high-valued targeted applications, for example, counter-terrorism or drug discovery, without having to overcome the full challenge of natural language understanding by machines. Though I seriously doubt that the full understanding problem will be achieved by

2020, I think it is highly likely that more focused applications of text mining will become commonplace in this timeframe.

With the rise of text mining, I see a coming intersection of two long distinct histories of computational use, one supporting organizations, the other supporting individuals. Enterprise data computing embodied by mainframes, relational databases, ERP 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, processing over language-based information will surpass the processing over operational data originally captured in structured databases.

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

References

1. Vannevar Bush, *As We May Think*, Atlantic Monthly, July 1945, http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm

2. Karen Sparck-Jones, Peter Willett, Readings in Information Retrieval, 1995, Morgan Kaufmann Publishers, San Francisco, CA.

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, http://searchenginewatch.com

5. Licklider, J.C.R., Man-Computer Symbiosis, IRE Transactions on Human Factors in Electronics, HFE-1 (March, 1960), 4-11, http://sloan.stanford.edu/mousesite/Secondary/Licklider.pdf

Bio

Ramana Rao is CTO 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 ten 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 http://www.ramanarao.com/informationflow/