Meeting of the MINDS:
An Information Retrieval Research Agenda

Jamie Callan, Carnegie Mellon University (Chair)
James Allan, University of Massachusetts, Amherst
Charles L. A. Clarke, University of Waterloo
Susan Dumais, Microsoft Research
David A. Evans, JustSystems Evans Research
Mark Sanderson, Sheffield University
ChengXiang Zhai, University of Illinois at Urbana-Champaign

This report is one of five reports that were based on the MINDS workshops, led by Donna Harman (NIST) and sponsored by Heather McCallum-Bayliss of the Disruptive Technology Office of the Office of the Director of National Intelligence's Office of Science and Technology (ODNI/ADDNI/S&T/DTO). To find the rest of the reports, and an executive overview, please see http://www.itl.nist.gov/iaui/894.02/minds.html.

1 Introduction

Since its inception in the late 1950s, the field of Information Retrieval (IR) has developed tools that help people find, organize, and analyze information. The key early influences on the field are well-known. Among them are H. P. Luhn's pioneering work, the development of the vector space retrieval model by Salton and his students, Cleverdon's development of the Cranfield experimental methodology, Spärck Jones' development of *idf*, and a series of probabilistic retrieval models by Robertson and Croft. Until the development of the WorldWideWeb (Web), IR was of greatest interest to professional information analysts such as librarians, intelligence analysts, the legal community, and the pharmaceutical industry.

In the early 1990s a combination of inexpensive disk storage and the development of the Web changed the field dramatically. Organizations of every size began to amass digital document collections, and Web search companies began making copies of the Web. As document collections became large, search engines quickly became the preferred method of finding information. Perhaps most importantly, the Web evolved from a tool for scientists to a communications medium for the public. Suddenly everyone was a user of search engines.

The IR field has evolved rapidly during the last decade. Among the many noteworthy developments are multilingual and cross-lingual search; broader recognition of the importance of multiple document representations; new retrieval models based on statistical language models; the use of machine learning to set model parameters; the reawakening of interest in question-answering; the open-source movement, which gives a broad research community access to free, high-quality search engines; and a great broadening of the IR research community. Among the less positive developments, the commercialization of web search has caused a significant shift in the balance of knowledge between industry and academia; large web search engines have Web data, user data, and computer hardware that researchers cannot begin to reproduce, raising concerns about the quality and relevance of some areas of academic research. The ways in which people produce, find, and use information are evolving rapidly, which means that

information retrieval is far from being a "solved problem." Indeed, the increasing complexity and scope of IR systems has greatly multiplied the number of open research questions.

As our group considered the state of IR research, there was much to be proud of, but also some concern. IR has a strong tradition of serious experimental evaluation, which has served the field well, but which also can encourage incremental research on old corpora and discourage research on new topics that are difficult to evaluate. This report briefly surveys recent influences on the IR field, and then turns its attention to some topics that we believe the field should be considering seriously.

2 Recent Influences

The growth of online information demands powerful information management tools to help people manage information effectively and efficiently. As new data characteristics and application needs emerge, they raise new challenges for IR research.

Globally, the Web is no doubt the biggest force driving recent information growth. The complete freedom and low-cost of publishing on the Web have several implications from the viewpoint of information management. Firstly, the amount of online information has grown and continues to grow rapidly, making it a significant challenge simply to handle the *scale* of the Web. For example, it is impossible to maintain a complete and current index of all of the information on the Web, making Web crawling a new challenge. Secondly, there is no control over the quality of information. Spelling errors are common, the validity of information may vary significantly depending on the information source and author, and spam pages are intentionally created for profit or deception. Modeling *information quality* has become an essential component of Web search. Thirdly, the data and content are highly heterogeneous along every dimension, for example language, genre, and degree of content structure. *Data heterogeneity* raises challenges for all of the components in an IR system. Finally, the *users are now everyone*. The great majority of users are ordinary people with little search expertise who benefit greatly from user-friendly search support. The search activity of this large, worldwide user population creates log data that offers unprecedented opportunities to better understand users and improve search tools.

The amount of digital *enterprise information* stored by companies, governments, and other organizations has also grown very rapidly during the last two decades. In order to maximize productivity, it is essential to manage all of an organization's information so that it can be easily published, organized, transformed, and consumed. *Information integration* is especially important. Many of the characteristics of the Web data discussed above are also shared by enterprise data, especially the heterogeneity of data and ordinary people as users. One additional characteristic in enterprise search is that there is often a *task environment or context* associated with search, which can be leveraged to better model a user and provide appropriate task support.

As personal computers become common household appliances, people begin accumulating information individually. With distributed service providers, our personal information is often scattering among family computers, work computers, and the Internet (e.g., multiple email and blog accounts), making it a challenge to manage personal information. Personal information management shares some of the characteristics of public Web information and enterprise

information, but also includes *privacy protection*, which is widely recognized, but not yet widely studied by IR research.

These influences have affected recent research in information retrieval, and have motivated new evaluation tasks in influential evaluation workshops. For example, in TREC, the Blog, Enterprise, Genomics, Legal, and Spam tracks focus research attention on documents of different types, structure, and quality, as well as search in Web, email, medical, and legal domains. The CLEF workshop facilitates cross-lingual and European language retrieval research; in 2007, there were eight tracks addressing genres such as news, Web, and scientific documents. The NTCIR workshop addresses Asian and cross-Asian language tasks in both retrieval and summarization. The INEX workshop studies retrieval of structured (e.g., XML) documents. In addition to these search-oriented evaluations, DUC moves towards task-support by studying summarization of information.

The IR research community emphasizes empirical evaluation, so these evaluation workshops play a major role in directing research. Many recent IR research publications used evaluation resources created by these workshops. However, the field has only scratched the surface of the challenges people and organizations face in finding, managing, and using information now that so much information is available. Many challenges, such as contextualized search, personal information management, information integration, and task support, have not yet been addressed through such workshops.

In the rest of this report, we further elaborate some of the major challenges and present corresponding research agendas.

3 Aren't Commercial Search Engines Enough?

The last decade has seen the emergence of large-scale commercial Web search engines (e.g., Google, Yahoo! and Microsoft). These search engines grew out of core research in information retrieval algorithms and systems, but also addressed the challenges of scale, combining link topology with content, and supporting navigational as well as informational tasks. These engines have transformed the way in which people find, publish and share information. For example, on any given day in September 2005, about 20% of Americans used a Web search engine.[1] In countries around the world, Web search engines have become one of the preferred methods of finding information.

Good search engines are also available, for free, for personal information stored on desktop computers.

Given the widespread availability of free text search and the highly competitive nature of the search business, is it still necessary for funding agencies to invest in information retrieval research? We believe that the answer is yes.

In spite of the tremendous success of Web search engines, Web search is still in its infancy. Many searches are unsuccessful (up to 50% of searches don't result in a single click), and even

[1] Pew Internet and American Life Project.

those that are successful are often harder than they should be. The tools provided are rudimentary – searchers specify their information needs by typing a few words into a small rectangle, the system returns a long list of search results, and the searcher either follows a link or tries again. There are many challenges in extending search as we know it today, only some of which commercial search engines can or will take on.

An individual's information landscape is rich, cluttered, and diverse, for example including email, information stored on desktop computers at home and work, information resources provided by an employer, subscription-based resources, and blogs. Web search remains important, but it is just one part of an increasingly heterogeneous information landscape (see Section 4). Today's search engines generally ignore the context in which information is used, in part because they know a lot about general population behavior, but very little about an individual, her history, or the context of her current search (see Section 5). An increasingly important challenge for the coming decade is organizing, managing, summarizing, and mining the information people find (see Section 6); few people are able to use the tools available for these tasks today. There is surprisingly little good scientific data about how people use the information tools available to them today, or where today's tools fail to meet their needs, because experimental research methodologies haven't kept up with the rapid pace of change (see Sections 7 and 8). Finally, in recent years *software applications* have emerged as a new class of users (see Section 9); neither commercial search engines nor today's research search engines are designed for this new class of user.

It would be foolish to expect commercial search engines to do all that needs to be done in the coming years; they are profit-driven companies with specific objectives. Society has a long history of conducting basic science and making it available for others to build upon. The scientific community's responsibility is to discover new knowledge about how people seek, use, and organize information, and to develop new tools that assist people in achieving their goals. Some of what the scientific community discovers will be immediately useful to today's commercial search engines; some of it may be the seeds of tomorrow's new companies.

4 Heterogeneous Data ("Everyday Data")

IR research has traditionally focused on well-edited text, such as newspaper articles, journal papers, and Web pages. While well-edited text may remain the central data type, many people use a richer variety of formats and media, including blogs, instant messaging, text messaging, email, speech, video and images. Their information is acquired from diverse sources, and varies widely in its level of quality and trustworthiness. Some sources of information—for example, email, the Web, and instant messaging—may include information from adversaries attempting to deceive, defraud, or otherwise cause harm. Software applications may automatically provide document structure, text-level annotations, or rich metadata.

This hodgepodge of data presents challenges and opportunities for IR systems. The volume and complexity of the data generated by these sources precludes any possibility of manual cleaning or organization. While an IR system might filter out material that is outright harmful or adversarial, such as spam or viruses, the remaining material must be retained and made available for searching and browsing. The challenge increases when information arises as a mixture of

data types. For example, a recorded meeting or presentation might consist of a mixture of audio, video, slides, and notes. A transcript might be extracted from the audio, text and images might be extracted from the slides, the material might be automatically annotated by natural language processing, or manual tagging may be added incrementally as users access the information at later times.

To cope with this mixture of data, IR systems must seamlessly integrate and correlate information across a variety of media, sources, and formats. The relationships between diverse elements must be apparent to the IR system, and must be exploited to improve information access. Each source and format cannot have its own interactive search interface. The IR system must seamlessly merge information and adapt its presentation as new sources become available, learning from implicit user feedback as appropriate.

Substantial research is required to create IR systems capable of the required flexibility and adaptability. New retrieval models incorporating multiple sources of evidence must be developed. Tools to properly triage and integrate information from diverse sources must be implemented. Finally, new evaluation methodologies are needed to measure the performance of these systems.

5 Heterogeneous Context

Today's retrieval systems are the same for everyone regardless of who they are, where they are, when they are searching, what they are searching for, and why they are searching. As information technologies are used by an increasingly diverse user population for increasingly diverse tasks (finding, learning, monitoring, communicating, planning) search technologies and interfaces will need to be extended and improved. Consider, for example, a teacher preparing a lesson plan on global warming. Web search engines currently provide them with very minimal search and information management tools – the teacher issues a query by typing into a tiny search box, evaluates a long list of results one at a time, cuts and pastes interesting possibilities to a document or slide deck, and then repeats this process, over and over again. Understanding, representing and exploiting contextual information can transform how people perform this and other information discovery, analysis and synthesis tasks.

While there are many different factors that characterize searchers and tasks, three important classes of contexts seem widely applicable. First, our tools must do a better job of *understanding the user* who is asking the question and the previous knowledge and skills that she brings to bear on the problem. It is important to note that by "user" we mean both individuals as well as larger social or organizational groups (e.g., doctors, grade-school students, mobile users with small devices) that might be satisfied by different kinds of information resources. Representing the user includes understanding both short-term characteristics, such as the previous queries they have issued or sites visited in the session, and longer-term characteristics and preferences, such as the kinds of information they create and read, or their expertise in the domain of interest. Characteristics of users could be modeled using a wide range of formalisms using information gathered either implicitly or explicitly. Second, the field needs to better understand and represent the underlying *information domains* (see also Heterogeneous Data, above). Instead of returning a long list of results, richer analyses of information sources would allow our tools to

show relationships across documents as well as richer entities within documents. Content, metadata and usage patterns are key sources of information that can be used to identify relationships. Finally, the *larger task* the user is trying to accomplish shapes both the kinds of information that is needed and how it should be presented. There are few tools to help people organize and digest information.

Search is not the end goal. It is a tool that can help people accomplish other tasks – certainly a very important part of the process, but only a part. The more our tools understand the context of the search – the user, the domains of interest and the large tasks – the better they can deliver the right information to the right people in the right way.

6 Beyond the Ranked List: Information Analysis & Organization

Traditional IR has emphasized matching query terms (often just "keywords") to document content, largely because it requires very little understanding of language meaning. Relevance ranking is a natural refinement to the simple process of returning documents in the order in which they are found; indeed, important and effective methods have been developed that make ranking a genuine "feature" of IR systems. Many ranking approaches make use of the distributional characteristics of terms in collections and local document contexts (e.g., idf x tf type term weighting schemes) to establish document scores. On the Web, such techniques have been augmented by methods that take advantage of the explicit cross-document citations (links) that characterize hypertexts (e.g., providing "authority" scores via algorithms such as PageRank) to further discriminate among documents. In all cases, the goal is to return to the user "the best documents first" – which is an important system function when the documents "hit" by query terms may number in the millions.

However, while some kind of ranking will no doubt be part of any modern IR system, the problems of heterogeneous data, scale, and non-traditional discourse types reflected in the documents, along with the fact that search engines will increasingly be integrated components of complex information management processes, not just stand-alone systems, demand new modes of system response to a query. When a person searches e-mail, she may not be interested in retrieving a set of ranked messages. Instead, she may want a minimal set that displays the history of a correspondence (not necessarily found in a "thread"). When a person looks for information on a health problem, she may not want expert-grade authoritative articles, but a few that are representative of the health-professional view. In short, as people move from finding documents to using them, and as the user model is enriched from a set of query terms to a rich context combining goals, work history, social relations, and more, relevance ranking is likely to become less dominant.

One of the problems of ranked lists is that they do not reveal relations that may exist among retrieved documents. Highly similar documents (possibly copies or version of the same document) may be retrieved as separate items; documents that are parts of a common source (e.g., a book) are disconnected in the ranked list; material that is critically ordered in time (e.g., email correspondence or software documentation notes) is presented without consideration of history. We can imagine techniques to address any one such problem – e.g., clustering at high threshold to remove or group redundant documents – but there is today no general methodology for

matching the system's response to the type of information in the response set or to the user's task or need.

To begin to address this problem, we believe the next generation of IR systems will have to provide specific tools for information transformation and user-information manipulation. Tools for information transformation in real time in response to a query will include, for example, (a) clustering of documents or document passages to identify both an information group and also the document or set of passages that is representative of the group; (b) linking retrieved items in timelines that reflect the precedence or pseudo-causal relations among related items; (c) highlighting the implicit social networks among the entities (individuals) in retrieved material; and (d) summarizing and arranging the responses in useful rhetorical presentations, such as giving the gist of the "for" vs. the "against" arguments in a set of responses on the question of whether surgery is recommended for very early-stage breast cancer. Tools for information manipulation will include, for example, interfaces that help a person visualize and explore the information that is thematically related to the query. In general, the system will have to support the user both actively, as when the user designates a specific information transformation (e.g., an arrangement of data along a timeline), and also passively, as when the system recognizes that the user is engaged in a particular task (e.g., writing a report on a competing business). The selection of information to retrieve, the organization of results, and how the results are displayed to the user all are part of the new model of relevance.

7 What Do People Really Do?

Many people have access to many search engines and many document collections, as well as many *types* of search engines and document collections. Many people also have access to text analysis and mining tools, primarily through Web sites or enterprise software. IR tools are now "everyday" tools for many people. In spite of this widespread use, surprisingly little is known about how most people use information retrieval tools "everyday", in part because much of the data they search or analyze is personal, private, or confidential. IR does not have well-developed methodologies for working with such data or producing reproducible research from it, so this type of research tends to be neglected. Instead, the field remains bound to research methodologies defined in the early 1970s.

Industry has learned to study how people search and interact with data in their daily lives. Web search engines, large digital libraries, and e-Commerce sites have well-developed methodologies for tracking users and their interactions with information resources. However, companies do not share such information easily, for competitive reasons and because of the same privacy issues that hamper the research community, so this knowledge is confined within individual companies. Although there is value to having such basic knowledge in the public domain, it will remain in the private sector until IR has "off-the-shelf" privacy preserving research methodologies.

A modern research methodology for information retrieval user studies would require pervasive awareness of, and techniques for, anonymization and privacy protection, perhaps developed in collaboration with researchers who have studied privacy issues in other fields (e.g., medicine). One approach is to develop methods that allow usage data to be shared; this may be a difficult goal to achieve. An alternate approach is to develop standard tools and methodologies for

capturing usage data, to make it easier to do this kind of research, and to make it more likely that experiments by different people are roughly comparable even if conducted on different users.

IR has been well-served by the Cranfield experimental methodology, which is based on sharable document collections, information needs, and relevance assessments. However, as the field develops the tools to study how users actually use information retrieval tools on a daily basis, it is likely that a broad reassessment of IR experimental evaluation will occur. Today's tools provide little guidance in how to evaluate the display of search results, document clustering, or other system components that people use every day. All areas of IR research can probably benefit from studying real users "in the wild".

8 Evaluation

Information Retrieval is an empirical science; the field cannot move forward unless there are means of evaluating the innovations devised by researchers. IR has been a leader in Computer Science in understanding the importance of evaluation and benchmarking; thanks to the efforts of academics in the US and UK, the field established large-scale shared evaluation platforms in the 1960s and 70s. The platforms inspired the modern evaluation campaigns of today: TREC, CLEF, NTCIR, INEX, etc. However the methodologies conceived in the early years of IR and used in the campaigns of today are starting to show their age and new research is required to understand how to overcome the emerging twin challenges of scale and diversity, and how to go beyond the well-established Cranfield experimental methodology.

Scale

The methodologies used to build test collections in the modern evaluation campaigns were originally conceived to work with collections of tens of thousands of documents. The methodologies were found to scale well, but potential flaws are starting to emerge as test collections grow beyond tens of millions of documents. Potential solutions are being investigated, but the long term stability of the test collections formed with the new approaches is still unclear. The large search engines have their own solutions, but their approaches are extremely costly. Support for continued research in this area is crucial if IR research is to continue to evaluate large scale search.

Diversity

With the rise of the large commercial Web search engines, some believed that all search problems could be solved with a single engine retrieving from a one vast data store. However, it is increasingly clear that evolution of retrieval is not towards a monolithic solution, but instead to a wide range of solutions tailored for different classes of information and different groups of users or organizations. A search on any popular Internet engine returns a Web search along with a range of alternate options for the user to explore: search of images, news, discussion groups, books, products, academic sources, etc. Each tailored system on offer requires a different mixture of component technologies combined in distinct ways and each solution requires evaluation. The growth of collection types potentially searchable shows no sign slowing. Diversity is particularly acute in enterprise search where vendors tell stories of having to re-tune their search engine for each organization they sell their software too. Each re-tune, demands a new test collection.

One might think that the plethora of worldwide evaluation campaigns (TREC, CLEF, NTCIR, INEX, etc) and their broad range of tracks could keep up with the growth in search applications. However, there are only between 100-200 public test collections available at the moment. This number only scratches the surface of what is required. Up until now test collections have been formed by large numbers of researcher groups working together to build a collection. In an environment of diversity, although there will still be search problems that many may wish to address collectively, there will be more search problems that are only tackled by individual researchers or research groups. To address the diversity problem, it will be necessary to determine reliable methods of test collection formation that can be conducted by single organizations or individuals. As with the problems of scale, research is starting to be conducted in this area, however, much is still to be done before a rigorous reliable methodology for testing is determined by IR researchers.

Beyond the Cranfield Methodology

We have argued in Sections 5 and 6 for the importance of understanding the context in which a search is performed, and of having tools that help a person organize and analyze retrieved information. How should such search improvements be evaluated? The use of context poses a severe challenge to the notion of a test collection, which will need to include rich user and task models. Evaluation of new tools will require development of new metrics and methodologies; the difficulty of evaluating clustering algorithms – an old and well-established research area that doesn't have correspondingly well-established metrics – reminds us that robust and well-accepted experimental methodologies are significant research accomplishments that do not happen often. If the past is any guide, a lack of well-accepted experimental methodologies will be a significant obstacle to making some of the improvements suggested in preceding sections.

9 IR in Service of Human Language Technology Applications

IR research during the last forty years has focused on finding, organizing, and summarizing information for use *by people*. However, during the last decade *software applications* have emerged as a new class of IR system users. Question answering and information distillation systems are examples of this class of applications. The typical architecture involves a text search engine to efficiently gather "raw" information from a text database, and more sophisticated or specialized processing on the returned documents. In much the same way that e-Commerce systems are built on relational database systems, these applications are built on search engines.

Although this emergence is underway, current text search engines are designed for the kinds of information needs that people have, not the kinds of information needs that software application has. In cross-area discussions among the MINDS groups, people from other research areas described different types of information needs. For example, a speech recognition system might form a language model of the last few minutes of speech (perhaps it is about basketball), pass it to a search engine as a "find similar documents" query, receive back a set of documents (about basketball), and use them to form a more accurate language model for predicting the speech that will be encountered next.[2] Machine translation and natural language processing researchers described similarly specialized needs, for example, support for recognizing viable translation

[2] Thanks to Sanjeev Khudanpur for this example.

hypotheses, constructing parallel/comparable corpora, and extending concept hierarchies. The common theme was that their research requires routine access to large text databases. Often there is a specific information need that requires quickly gathering a "small" amount of text – just what full-text search engines are designed to do.

Recognizing and providing strong support for search by software applications is important because much recent human language technologies research is now data-driven. NLP, MT, and speech researchers need routine access to large corpora. Currently they must build specialized "one off" solutions to obtain such access; what they need is text search engines that support their information needs.

There is little systematic study of the information needs of typical software applications across human language technology, and how to support them well, but one might make some guesses about some of the required elements. Many of the required elements are rooted in longstanding IR theory, if not so much IR practice. Search engines may need to routinely support multiple representations of a text, for example, a text-based representation, (hierarchical) text annotations, metadata, and/or controlled vocabulary representations. Retrieval models and query languages developed for structured (e.g., XML) documents might be extended to support multiple, loosely-synchronized representations. Greater study of indexing methods and optimization may be required to provide efficient support for some types of information needs.

The transition to supporting the information needs of software applications is already underway, but it proceeds haphazardly, and with little recognition by the field. Greater recognition by both researchers and funding agencies would accelerate this trend, eventually producing a new generation of search engines that would better serve both interactive users and a broad set of language-based software applications.

10 A Common Core of Research

As described in the preceding sections, we believe that the IR research community should focus on several key challenges, among them heterogeneous content and context, information analysis and organization, and new evaluation paradigms. With a few notable exceptions, IR research collections are limited to data of a single kind, for example collections of newswire, blogs, recognized spoken news, or email messages; a key challenge is constructing and using massive collections of data of varying genre, different topics, and numerous formats. The quintessential IR problem has been search, but once information is found, it is just as important to impose or find structure in the midst of a mountain of data; progress on this problem is hampered by insufficient models of long-term user behavior and preferences, by a limited understanding of how data is usually organized, and by the lack of theories and techniques for acquiring that information. Finally, the Cranfield evaluation methodology has been the foundation of substantial advances, however it does not easily extend to the highly dynamic and nearly ubiquitous way in which IR is now used, or many problems that are becoming important; new experimental methodologies are required.

These challenges encourage researchers to take a broader view of information, users, and tasks, which we believe will lead to significant improvements in IR techniques.

It would be unfortunate, however, if our enthusiasm for new research directions diverted attention from fundamental IR research. Information retrieval has a long history of research on document representations, formal retrieval models, and evaluation that provides solid foundations upon which other research is built. Without such foundations, individual research progress is slower, and research by different individuals is difficult to compare. These foundations will need to be extended, for example, to include long-term user modeling and a more systematic approach to tailoring or training general solutions for specific tasks, users, or datasets. Research progress on "core" IR topics is of benefit to the field as a whole, and must not be neglected.

Each of the challenges described above will depend on advances in a number of areas. How can data be represented so that it can be compared across media, across formats, or in varying contexts? How can retrieval and organization processes be best modeled with such data? What does it mean to model a person's long-term search and organization needs rather than just one-shot querying? How can existing and new core IR techniques be understood and developed so that they are broadly applicable across a range of IR tasks, data, and systems? Each of those questions has been explored in the past, but the expanding role of information in everyone's lives, the need for more effective and appropriate search and organization to support work and recreation, and the potential for embarrassment or even disaster when information is missed – all of those highlight the need for innovative and successful work in Information Retrieval.