Salton Award Lecture

Information Retrieval and Computer Science: An Evolving Relationship

W. Bruce Croft
University of Massachusetts Amherst
Department of Computer Science
Amherst, MA 01003-4610

ABSTRACT

Following the tradition of these acceptance talks, I will be giving my thoughts on where our field is going. Any discussion of the future of information retrieval (IR) research, however, needs to be placed in the context of its history and relationship to other fields. Although IR has had a very strong relationship with library and information science, its relationship to computer science (CS) and its relative standing as a sub-discipline of CS has been more dynamic. IR is quite an old field, and when a number of CS departments were forming in the 60s, it was not uncommon for a faculty member to be pursuing research related to IR. Early ACM curriculum recommendations for CS contained courses on information retrieval, and encyclopedias described IR and database systems as different aspects of the same field.

By the 70s, there were only a few IR researchers in CS departments in the U.S., database systems was a separate (and thriving) field, and many felt that IR had stagnated and was largely irrelevant. The truth, in fact, was far from that. The IR research community was a small, but dedicated, group of researchers in the U.S. and Europe who were motivated by a desire to understand the process of information retrieval and to build systems that would help people find the right information in text databases. This was (and is) a hard goal and led to different evaluation metrics and methodologies than the database community. Progress in the field was hampered by a lack of large-scale testbeds and tests were limited to databases containing at most a few hundred document abstracts.

In the 80s AI boom, IR was still not a mainstream area, despite its focus on a human task involving natural language. IR focused on a statistical approach to language rather than the much more popular knowledge-based approach. The fact that IR conferences mix papers on effectiveness as measured by human judgments with papers measuring performance of file organizations for large-scale systems has meant that IR has always been difficult to classify into simple categories such as "systems" or "AI" that are often used in CS departments.

Since the early 90s, just about everything has changed. Large, full-text databases were finally made available for experimentation through DARPA funding and TREC. This has had an enormous positive impact on the quantity and quality of IR research. The advent of the Web search engine has validated the longstanding claims made by IR researchers that simple queries and ranking were the right techniques for information access in a largely unstructured information world. What has not changed is that there are still relatively few IR researchers in CS departments. There are, however, many more people in CS departments doing IR-related research, which is just about the same thing. Conferences in databases, machine learning, computational linguistics, and data mining publish a number of IR papers done by people who would not primarily consider themselves as IR researchers.

Given that there is an increasing diffusion of IR ideas into the CS community, it is worth stating what IR, as a field of CS, has accomplished:

- Search engines have become the infrastructure for much of information access in our society. IR has provided the basic research on the algorithms and data structures for these engines, and continues to develop new capabilities such as cross-lingual search, distributed search, question answering, and topic detection and tracking.
- IR championed the statistical approach to language long before it was accepted by other researchers working on language technologies. Statistical NLP is now mainstream and results from that field are being used to improve IR systems (in question answering, for example).
- IR focused on evaluation as a research area, and developed an evaluation methodology based on large, standardized testbeds and comparison with human judgments that has been adopted by researchers in a number of other language technology areas.
- IR, because of its focus on measuring success based on human judgments, has always acknowledged the importance of the user and interaction as a part of information access. This led to a number of contributions to the design of query and search interfaces and learning techniques based on user feedback.

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.
ACM 1-58113-646-3/03/0007.

Although these achievements are important, the long-term goals of the IR field have not yet been met. What are those goals? One possibility that is often mentioned is the MEMEX of Vannevar Bush [1]. Another, more recent, statement of long-term challenges was made in the report of the IR Challenges Workshop [2]:

- *Global Information Access*:
  Satisfy human information needs through natural, efficient interaction with an automated system that leverages world-wide structured and unstructured data in any language.

- *Contextual Retrieval*:
  Combine search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information need.

These goals are, in fact, very similar to long-term challenges coming out of other CS fields. For example, Jim Gray, a Turing Award winner from the database area, mentioned in his address a personal and world MEMEX as long-term goals for his field and CS in general [3]. IR's long-term goals are clearly important long-term goals for the whole of CS, and achieving those goals will involve everyone interested in the general area of information management and retrieval. Rather than talking about what IR can do *in isolation* to progress towards its goals, I would prefer to talk about what IR can do *in collaboration* with other areas.

There are many examples of potential collaborative research areas. Collaborations with researchers from the NLP and information extraction communities have been developing for some time in order to study topics such as advanced question answering. On the other hand, not enough has been done to work with the database community to develop probabilistic retrieval models for unstructured, semi-structured, and structured data. There have been a number of attempts to combine IR and database functionality, none of which has been particularly successful. Most recently, some groups have been working on combining IR search with XML documents, but what is needed is a comprehensive examination of the issues and problems by teams from both areas working together, and the creation of new testbeds that can be used to evaluate proposed models. The time is right for such collaborations.

Another example of where database, IR, and networking people can work together is in the development of distributed, heterogeneous information systems. This requires significant new research in areas like peer-to-peer architectures, semantic heterogeneity, automatic metadata generation, and retrieval models.

If the information systems described above are extended to include new data types such as video, images, sound, and the whole range of scientific data (such as from the biosciences, geoscience, and astronomy), then a broad range of new challenges are added that need to be tackled in collaboration with people who know about these types of data.

There should also be more cooperation between the data mining, IR, and summarization communities to tackle the core problem of defining what is new and interesting in streams of data.

These and other similar collaborations will the basis for the future development of the IR field. We will continue to work on research problems that specifically interest us, but this research will increasingly be in the context of larger efforts. IR concepts and IR research will be an important part of the evolving mix of CS expertise that will be used to solve the "grand" challenges.

ACKNOWLEDGMENT

I want to thank Keith Van Rijsbergen, my M.Sc. advisor in Australia, my Ph.D. advisor in England, and my friend, for showing me what an exciting area IR can be, and for inspiring me by his example to combine modeling and experimentation to understand the retrieval process.

REFERENCES

[1] Bush, V. "As we may think", *Atlantic Monthly*, 176, 1945, p. 101-108.

[2] Allan, J. et al, "Challenges in Information Retrieval and Language Modeling", *SIGIR Forum* (to appear).

[3] Gray, J. "What next? A dozen information technology research goals", *JACM*, 50, 41-57 (2003).