Information Retrieval—A View of its Past, Present, and Future

Charles T. Meadow
Faculty of Information Studies, University of Toronto
140 St. George Street, Toronto ON M5S 3G6
meadow@fis.utoronto.ca

Abstract

The development of information retrieval systems is reviewed from its early history to the present time. Emphasis is placed on user languages and functionality and the effect of available technology on these aspects. Through most of its history, retrieval has depended on the use of surrogate records because technology did not permit direct searching of documents. The languages used for querying systems were formal and mathematical, to the advantage of those writing programs but not those using the systems.

Today, distance from user to computer is not a factor in cost, and the time need for browsing is only a minor one. Query languages typically allow use of natural language and retrieved records are ranked by similarity to queries. New methods of query enhancement—finding related terms not explicitly included by the user—are coming into use. Many users are unable to evaluate retrieved items, hence accept almost anything that "looks good."

Future systems must be able to interpret user queries in context, assisting users to understand what was retrieved and how a query might be restated to improve results. Direct search of graphic images will someday be common. As systems become more sophisticated and more widely used, it will be essential that users become better trained in understanding how they work and how to evaluate output.

Introduction

The purposes of this paper are to review the historical development of mechanical aids to information retrieval, identify current trends, and indicate what seems still to be needed. It will necessarily be a personalized view.

IR Before 1950: The Surrogate Record

From earliest times information retrieval has been based on classifying or indexing, i.e., creating and searching of a surrogate for a document. As a practical matter, little else could have been done, but these systems, if used today, would be highly unsatisfactory. They depended on creation of a surrogate record and on the knowledge on the part of the user of what was likely to be in the surrogate of a document of interest.

Mechanical aids to information retrieval first appeared when the first library catalogs were created, some say in Sumer four millennia ago [Kramer, 1963], some say in Greece merely two millennia ago. [Johnson, 1970] These catalogs were basically lists of book titles or broad subject categories. It took until the mid-19th century for the card catalog to be developed, which allowed for more than one entry point or search key. It was easy to update but, of course, there would normally be only one copy of it; the user had to come to the central location. As libraries grew, such catalogs became less useful because the same limited number of access points were used for ever larger collections, but little could be done to accomplish significant improvements for lack of technology. Further, end users were rarely well trained in use of the tools. Book catalogs allowed for multiple copies but updating them created problems.

1950s: Faster Searching of Surrogates

The decade of the 1950s saw a great many changes in the way documents were indexed and searching was done. Why at that particular time? There was a push-pull situation. The push came from the fact that during World War II a great deal of secret scientific work had been done. After the war there was a large backlog of material to be made public. The pull was a combination of a greatly increased interest in science, caused by the war just ended, and the prospect of a future war that would clearly involve new technologies if it came about.

Probably the most frequently cited work in the field was the well known "As We May Think," by Vannevar Bush, [Bush, 1945] who had been science advisor to U.S. President Roosevelt during World War II. We have still not fully achieved his vision, although in some ways we have surpassed it. His idea was of an interactive machine with a huge memory that could find information on request, store the user's comments or notes in the database, and make associations among works stored. The storage technology was microfilm. This, in concept, was very much like what some people think the World Wide Web is today. But the web tends to lack that intellectual component.

In the 1950s, we began to see primitive approaches to Bush's dream. Taube's *uniterm* method of indexing [Taube, 1953-65] allowed for far more descriptors per document than had previously been common at least with books. Searching was not mechanized but was enhanced by design of a card representing the term rather than the document. This was an elegantly simple solution, highly useful for modest size collections, but impossible for a collection like that of the Library of Congress or even my university's library of over eight million volumes. A variation on this system involved a large sheet of metal or stiff paper in which a hole was drilled at a point that represented a term and the coordinates of a hole drilled in the card represented a document number. Searching required placing two or more term cards on a light table. Any points of light that were visible would indicate by their location the numbers of documents having all the terms of interest. [Wildhack & Stern, 1958] The magnitude of a library that could be handled by such methods depended on the size of the sheet. There were obvious limits.

Another development of the 50s was Western Reserve University's special-purpose computer called the *Searching Selector.* A big advance, but its weakness was that it required indexing or cataloguing in such detail that few could master it, and if searchers could not anticipate what catalogers had done, there would be no search success. That point remains a key one in IR, even today. *Searchers must be able to anticipate how catalogers have described a subject in the surrogate.* Of course, they must also anticipate how the author described the subject in the original text, but the surrogate adds an extra step to the problem.

There were several interesting experiments in this period involving general purpose computers, mostly for military projects never made public, but there were no great break-throughs. Among the reasons: (1) It still took a long time to get a question answered and, if the results were not satisfactory, one had to try again. A "long time" could be a day, or at least a few hours because computers were not yet interactive and "jobs" had to be carried to a computer centre, and the results then carried back to the requester — no automatic transmission. (2) Computer memories were still not big enough to hold the kinds of databases that by then already existed in paper form.

These early systems established the principles of what a query language should be like. They were all quite similar. One specified an attribute of a record and a value for it and all records having that attribute value constituted a subset of the database. Boolean algebra was used to combine subsets in various ways. These languages were easy for mathematically-oriented people to understand and it was relatively easy to write programs to interpret them. Retrieved records would be presented in the order in which they occurred in the database, or sorted by some record attribute, such as *date*. As new systems were designed, we tended to stay with this form of query language as if its superiority over any other were clearly established.

1960s: Interaction with Search Computers

In the 1960s general purpose computers made three great strides: bigger memory, interactive access, and the United States Government-developed ARPANet, predecessor of the Internet. These meant that records and files could be bigger, that users could try something out, see results, and change their approach, all in a few minutes. It also meant that one could use a computer in another building, city, or country, with distance between user and computer having virtually no effect on the cost of usage.

1970s: Beginnings of Text Search

In the early 1970s there was almost an explosion of online, interactive information retrieval systems: Dialog, Orbit, Medline, and Lexis in the United States; CANOLE in Canada; and the European Space Agency's Information Retrieval Service (ESA-IRS). The average cost to use Dialog was about $US25 per hour, communications included; Lexis cost more; Medline was free to the medical world. I used to remind my students that $25 per hour amounted to about $8 per search and that was about the price of dinner in a restaurant (then), within the occasional reach of most students.

Typically, the languages of the day were still those of the 1960s. Boolean logic predominated, but at least the cost of browsing was getting less because of reduced communication costs and higher speeds. Quick browsing was able to overcome many weaknesses in the logic of queries.

Once these systems were developed, we still had several problems:

1. *Memory*. In the early 1970s, although computer memories were doubling in capacity about every two years, we still could not handle full-text in most systems. A few databases were approaching one million records, typically of about 1,000 bytes each. There might have been abstracts in bibliographic records, but not the full text of articles. The cost of printing large numbers of records locally for browsing was coming down but was still too high for much comfortable browsing.
2. *Indexing and Language*. If we could not store full text, we still had to rely on a surrogate. There could be more descriptors than for a conventional library card, but it was still a surrogate, not the actual text. As long as what could be searched was limited, the language for expressing what was sought was equally limited. I used to feel in those days that information retrieval had fallen far behind other major computer applications in developing user languages. For mathematics we had FORTRAN and BASIC, for business we had spread sheets and database systems such as Lotus 1,2,3. For operational IR systems, even through the 1970s, we still had only very primitive, mathematical IR query languages.
3. *User training*. Searchers through the '70s tended to only be professional librar-

ians or others specially trained. Yet, we were beginning to advertise to the world that these systems were for everyone. Users were often woefully ignorant, sometimes not recognizing the difference between a bibliographic record and the original document, not understanding basic Boolean algebra, or not being able to judge the value of documents or understand what the retrieved documents told them about how to revise a search query. Developers of IR systems generally acted as if they did not believe that end users, rather than professional librarians, would become the dominant group of searchers, and did not tend to adjust their systems accordingly. [Meadow, 1979]

1980s: Full Text, the Internet, and Search Engines

In the 1980s memories got still larger, full text records became the norm, and communications became ever more reliable and cheaper. All the problems remained, but businesses selling IR services became better established, meaning profitable. Universities were connected to search services in ever greater numbers. Law and medical schools in the United Sates and Canada could not exist if not connected to the principal retrieval services of their profession. One science information service acknowledged that its traditional market was saturated, that it already had every library in its subject field as customers. Growth now meant reaching out to individuals rather than only institutions. Computer science began to be seriously interested in information retrieval, perhaps evidenced by the first occurrence (that I could find) in print of the expression *search engine*, in 1984. [Holmes, 1984]

During this period, a few researchers, most notably Gerard Salton, [Salton, 1971, 1983, Salton & McGill 1983] were developing new ideas on how to search larger, complex databases with full text. The industry paid little attention at the time. One of the key areas of development was that of applying weights to documents or query terms, enabling records to be ranked according to their similarity to a query and also enabling users' ratings of retrieved records to be used to change the weights of terms in a subsequent query. While there are many ways to attack these problems, finding the "best" way to do this still escapes us, as does a definition of *best*.

Near the end of this decade there were many new networks in existence for connecting users with remote computers. The Internet, a new concept for a standard way of interconnecting networks, was developed allowing virtually any user to connect with any computer on any of these many networks. Its impact was to become profound.

1990s: The World Wide Web—Promises and Pitfalls

In 1991 the World Wide Web appeared. [Berners-Lee, 1996] This was a procedure for placing documents in storage in one computer, that would be available from any other that had Internet connections and documents could have links within them to other documents on the web. The growth of the 'Net and the 'Web were phenomenal and still are. They have changed, probably forever, some aspects of information retrieval, but left some problems still unsolved.

Accomplishments in this decade:

1. A vast number of documents is now available at what appears to be free of cost to users. Of course, access is not really free, but the costs are well hidden for most of us.
2. Graphics and sound recordings are now readily available as parts of retrieved records. These were rarely available before.
3. Search languages are improving.
4. Information retrieval now actually *is* an

activity for the average end user, without special training.

Problems:

1. There is, in general, *no* screening or filtering of documents that go onto the Web. This leads to moral, legal, and political problems and also technical problems of how to find the best information on a given topic.
2. The number of documents available makes the need for careful searching and evaluation more important than ever. Yet, the average user is much less qualified than was true two decades ago when searchers were professionals.
3. There are still the problems of how searchers must decide how to ask for a given subject and how they can learn to search and evaluate. Most of us have had the experience of asking on the web what seemed a well-structured query, only to retrieve over a million hits.

We have known for some time that all search problems are magnified as the size of the database to be searched increases. Any language ambiguities grow worse, never lessen, with size. Apparently, free browsing helps, but users must understand how to do this, what to look for, and what to do with information they can glean from a first try at retrieval.

Having graphics available is very valuable. Before we had this, an article might refer to a diagram, but the searcher could not see it without getting the original document, which could take days. However, does the unquestioned good looks of web documents make them more attractive than their content justifies? That gets back to the question of the user's ability to evaluate.

Related to evaluation, most of us have grown up learning that certain authors and publishers seem to produce the quality and selection of material we want, and others do not. This applies to newspapers and journals as well as to books. Libraries and museums are other traditional quality filters. With the web, the "publisher" is often someone or some organization about whom we know nothing.

There is no requirement to become a web user beyond having access to a computer connected to the Internet. Are the many untrained users getting what they need? Katzer [1998] thinks not, or worries that while we expect that the average school student might accept nearly anything on his or her subject, business and government managers may also be doing the same thing. Are they patient and skilful enough to demand the best information and are they able to find what they really want and recognize what they found?

While there remain problems, there have been major advances. Search languages are improving. Ranking of output is now almost universally available. If there is a problem with ranking, it is that users are not normally told the basis for it, hence the order is not necessarily what a knowledgeable user would want. But, it is far better than nothing.

Natural language queries are becoming common. While potentially an improvement over the old Boolean method, systems do not normally explain how these queries are interpreted. They normally do not deal with syntax or context. Users should understand this, but often do not realize that the machine tends to see only a list of words, unconnected by syntax.

Two new systems I have seen demonstrated, but are not yet widely available, have some very nice features. In one, a web search begins with a preliminary search of a modest number of reference works, such as encyclopaedias or dictionaries. From these, words associated with query words are recovered. "Associated" means statistical association, i.e., an associated word is one that co-occurs in the database with a query word at a certain frequency level and proximity. Also, words are stemmed allowing association to be based on common roots or looked up in dictionaries to find formally defined synonyms. In this way, an initial query is improved by being made broader. The revised query is then submitted to several web search engines—the choice being controlled by the user, as may the reference works used—and the first *n* documents from each search engine are merged into a single list and ranked. The result will not be perfect, but it should be better than what we have had before.

What Is Left to be Done?

Understanding seems to me the key issue—getting the machine to understand or simulate understanding of what a user is asking and helping users to understand what the retrieval system has done or is offering as a response to a query. And then, we must sooner or later get to the point where we can query graphics and sound recordings.

1. *Context.* An information retrieval system should be able to base word association or document matching techniques on a selected context. This, of course, would require a great many specialized dictionaries and thesauri, and possibly the searcher telling the system what that context is. We have made significant improvements in recent years, but far more could be done.

2. *Syntax.* We have made relatively little progress in interpreting the syntax of a natural language query. When one word modifies another syntactically, it changes its meaning and that should, in turn, change what the IR system is searching for. We still cannot handle *not* logic in natural language, as it is even difficult to do so in English. Finally, there is the anaphora problem—recognizing the intended meaning of words that have no inherent meaning, but refer to other words—such as *it* or *that*.

3. *Set Analysis.* I have mentioned the difficulty of users not being able to evaluate retrieved information in terms of its value, and to extract information from such records to be used in modifying a query. One way to help is to offer an analysis of a set of retrieved records. What terms occur often? What query terms never occurred? What terms tend to occur, elsewhere in the database, together with query terms? ESA-IRS had a version of this, called their *zoom* command, in the 1980s but it never became popular with other services.

4. *User Training.* I have also mentioned repeatedly that end users often lack the skills needed to compose queries and evaluate results. I believe that searching for information is a topic that should be taught in schools, beginning at the primary school level. I do not mean doing information retrieval on a computer, I mean the concept of information searching, whether in asking directions on the road, helping a customer in a shop, or searching for information in a library or database. It is too important to be left for later years. But, it should also be part of university curricula in specialized contexts. For that matter, it is just as important for potential doctors, lawyers, librarians, nurses, or salespeople.

5. *Searching Graphic and Sound Records.* Today, we cannot search a file of graphic images by drawing a picture of what we want to retrieve and having the system find matches, except in a few experimental systems. In general, graphic images must be converted to, or indexed as, text, codes, or numbers. We match fingerprint images, for example, not by directly comparing images, but by comparing versions of the images converted to codes and numbers. We shall surely want full graphic search capability in the future, as well as sound recordings, both without having had to first represent their content as text.

Conclusion

We have come a long way. Once not given

full priority by the computer science world, information retrieval is now so widely used by scientists, lawyers, managers, and people in many other professions that both its scientific and commercial importance are well recognized. We are not done with looking for better ways to do it.

References

Berners-Lee, Tim. WWW: past, present and future. *Computer,* 29(10) 1996, 69-77

Bush, Vannevar. As we may think. *Atlantic Monthly,* 176(1) 1945, 101-108.

Holmes, D. IPL funding new Mimer relational database for VME. *Datalink*, April 30, 1984, p. 1.

Katzer, Jeffrey. Personal communication, 1998.

Kramer, Samuel Noah. *The Sumerians, their History, Culture, and Character.* Chicago: University of Chicago Press, 1963, p. 226.

Johnson, Elmer D. *History of Libraries in the Western World.* 2nd edition. Metuchen NJ: Scarecrow Press, 1970, p. 57.

Meadow, Charles T. The computer as a search intermediary. *Online*, 3(3) July 1979, 54-59.

Salton, Gerard. The SMART Retrieval System — Experiments in Automatic Document Processing. Englewood Cliffs NJ: Prentice-Hall, Inc., 1971.

Salton, Gerard. *Automatic Text Processing.* Reading MA: Addison-Wesley, 1989.

Salton, Gerard; McGill, Michael J. *An Introduction to Modern Information Retrieval.* New York: McGraw-Hill, 1983.

Taube, M. ed., *Studies in Coordinate Indexing.* Washington: Documentation, Inc., 1953-1965, 6 vols.

Wildhack, W.A. and Stern, Joshua. The Peel-a-Boo System — optical coincidence subject cards in information searching. In Casey, Robert S.; Perry, James W.; Berry, Madeline M.; Kent, Allen, eds., *Punched Cards, Their Applications to Science and Industry.* New York: Reinhold, 1958, 125-151.