



## Special Section

### Text Retrieval Online: Historical Perspective on Web Search Engines

by **Trudi Bellardo Hahn**

---

The first online text retrieval systems appeared more than 30 years ago, and in the years since, they have continued to evolve. At a quick glance, it might appear that their development has been one long line of continuous improvement in functionality and usability. But is that true? Are all users - not just expert searchers - able to search with increasing facility and speed and with greater satisfaction with their results than the users of the early online era?

A part of the answer to that question can be revealed by examining the basic features of online systems, tracing when those features were introduced and comparing them with the functionalities and power of modern Web-based search engines. These features include search capabilities, browse capabilities and other miscellaneous helpful functions.

In the research that Charles Bourne and I are conducting for our forthcoming book on the early development of online systems, we are attempting to establish the milestones of invention and implementation. We will report on many details of development and implementation, in order to give credit to the many individuals who contributed to the progress of online retrieval. However, verifying dates, establishing priority of discovery and invention and giving proper credit to genuine trailblazers is a difficult and complex task. In this brief article, the examples given are not necessarily the very first or only of their type; they were, however, among the first solid trailblazing efforts that resulted in working systems.

What are the basic capabilities of online systems and search engines, when were they first introduced and by whom? Note that persons named are the individuals most often associated with a project. However, the pioneers did not work alone; they always were part of a research and development team. It is worth noting also that many of these search features had been conceptualized a decade or so earlier and used in serial searching of databases on magnetic tape.

#### Search Capabilities

The goal of a search capability is to match a user's specified information need with items in a database that will answer it. Two types of search capabilities are used: those that help to specify the relationship between terms in a search statement and those that facilitate the interpretation of a particular word.

#### Relationships Between Terms

- Boolean operations (AND, OR, NOT) enable set intersection, set union or set differences. At MIT in 1964, Mike Kessler developed TIP, which may have been the first online system to use Boolean logic in searching for bibliographic records.
- Proximity operations (NEAR) allow two or more words to be searched for as a single unit or phrase. Similar logical operations allow a user to restrict the distance allowed between two search terms within a record. The assumption behind these functions is that the closer two terms are found in a text, the more likely they are related to the same search concept. In 1969, at the Data Corporation in Dayton, Ohio, Richard Giering's Data Central system (the precursor of LEXIS) introduced full-text searching with word proximity capability allowed in the search formulation.
- Natural language query creation permits entry of a prose statement or question that describes the information the

user wants to find. In 1965 at System Development Corporation (SDC) in Santa Monica, California, Robert Simmons and Lauren Doyle conducted experiments with a system called Protosynthex. A small prototype full-text database of chapters from a child's encyclopedia was loaded on the system. Protosynthex could respond to simple questions in English with an "answer." It is not clear whether the system was truly online, but it was a pioneering effort in the use of natural language for text retrieval.

## Interpretation of a Particular Word

- Masking of characters in suffixes or prefixes, also called truncation or stem searching, allows searching and retrieval on part of a word. In 1964, the TIP system offered stem searching, including right- and left-hand truncation. In 1965, the capability of embedding a "wild card" in the middle of a word (for example, wom#n) was added to TIP.
- Numeric and date ranging permits specification of such data elements as date of publication. In 1967, the Data Central system introduced an arithmetic search capability to allow searching of numeric data fields.
- Search term weighting enables a user to indicate the importance of a term in a search statement. In 1965, at SDC, Herbert Isaacs and Jules Farrell conducted experiments with a system called TEXTIR (TEXT Indexing and Retrieval). The database consisted of robbery reports from the Los Angeles Police Department. A user could add a + sign to a search term to give it more weight.
- Elimination of stop words speeds retrieval by making some words not searchable. In 1963, at Stanford Research Institute, Douglas Engelbart and Charles Bourne used a large number of stop words in the SRI Online System. This was an important feature in the 1960s because of severe constraints on the amount of storage space available for the file indexes.
- Automatic incorporation of synonyms into search formulations improves recall without a user having to think of all the synonyms. This feature was built into the TEXTIR system in 1965.
- The capability of limiting a search to a specific field (author, geographic term, descriptor, title or other) improves precision. Field specification to retrieve records on the basis of bibliographic elements such as author, publication date and title words was an option in the SRI Online System in 1963.
- Case sensitivity, or distinguishing upper and lowercase, improves precision (for example, Carpenter vs. carpenter). In 1972, at IBM, Steven Furth introduced case sensitivity in a system called STAIRS. Obviously, it was important to distinguish "STAIRS" from "stairs."
- Citation searching allows retrieval of citations on the basis of cited references, thus bypassing the usual linguistic problems inherent in search terms. Citation searching is now a standard approach for such databases as Science Citation Index and Social Science Citation Index. It was first introduced in 1964 in the TIP system.
- The ability to request the system to find "more like this" permits a user who has retrieved at least one relevant record to identify similar documents without having to enter a new search strategy. "Find more like this" was not a feature of early systems until Ira Yermish experimented with it in the mid-70s at the University of Pennsylvania. In his system, a user could define "likeness" by various factors, including citations received, citations given, title words, authors, organizations, or any combination of factors.
- Fuzzy search provides the ability to locate words and concepts similar to the entered search term and is used to compensate for errors or variations in spellings of words. Based on the concept of fuzzy logic, in which membership in a set is not a simple yes or no matter, fuzzy search retrieves on the basis of a complicated algorithm of term weights and then assigns degrees of probability that an item matches the search topic.
- Although the concepts of fuzzy logic and fuzzy sets were first described in the 1960s by Lofti Zadeh, they were not applied to online text retrieval in the early online period.

## Browse Capabilities

Once a search is completed, browse capabilities help a user to determine which items are of interest and to select them to be displayed more fully. Since searches usually retrieve non-relevant items, browse capabilities assist in focusing on items that have the highest likelihood of meeting the information need.

- Ranking and relevance feedback bases the output order on predicted relevance values (a score intended to reflect the likelihood of a retrieved item to be relevant to a user's request). In 1964, the TEXTIR system computed a

- query relevance score for output records and generated records in rank order. In 1965, at SDC, J. L. Smith experimented with a system called MICRO that ranked output of bibliographic records from a database of 3,000 references to foreign-language science and technology journal articles. Also in 1965, at Lehigh University, Donald Hillman's LEADER system ranked output in order of relevance, as judged by the system.
- Zoning is the capability to display a key portion of each record retrieved, usually that will fit within a screen, and that is enough for a user to make a quick relevance decision. Zoning was available in 1963 in the SRI Online System.
  - Highlighting is the capability of displaying the words in retrieved records that match the search terms, as well as some surrounding text. Highlighting gives a user an indication of why an item was selected and whether the context of the retrieved record matches the user's need. In 1970, the Data Central system highlighted the matching terms in each retrieved record. Mead Data Central experimented with ways to achieve the highlighted effect. For black-and-white terminals, the words could be accented by blinking them, dropping them a little below the level of the others on the line, varying the light intensity, underlining them or flagging them with an arrow or asterisk or other special characters. They also developed contrasting colors for color monitors.
  - "How many records do you wish to see?" is a feature that attempts to accommodate each user's personal preferences for the amount of information desired on a topic. This capability was offered in 1965 in TEXTIR. It was an extremely useful feature of the early online systems, because records took a long time to print.

## Miscellaneous Capabilities

From the 1960s forward, online systems have offered additional functions that reduce the time and facilitate a user's ability to input queries and reduce the likelihood of entering a poor query.

- Iterative search is the capability to further modify the results from a previous search. It involves getting intermediate results and then being able to refine the search. In 1965, at SDC, Harold Borko experimented with a system called BOLD, which gave users intermediate results by providing posting counts of the number of records associated with each search term. In 1966, at Lockheed Missiles and Space Company, Palo Alto, California, Roger Summit's DIALOG system created numbered sets that could be further manipulated by entering only the set numbers.
- Canned or stored query is the capability to name a query and store it to be retrieved, executed and modified during a later session. In 1965, the TIP system permitted a stored search formulation to be recalled and incorporated into another search at a later time.
- Vocabulary browse allows a user to display in alphabetical order the words from the document database, beginning with any term, even a term fragment. Concept hierarchies and thesaurus expansion permit users to display hierarchical or conceptual relationships among terms in order to facilitate the selection of related terms. The display may be based on predetermined subject relationships or on statistical relationships (frequency of co-occurrence within same record or document, or on frequency of earlier users pairing terms in search statements). In 1966, the BOLD system enabled users to display an online thesaurus as a search aid. Also in 1966, at the University of Pennsylvania, Noah Prywes experimented with a system called Multilist, which permitted users to retrieve automatically all the records hierarchically subordinate or superior to a given term.
- Delivery of full source documents was a feature missing from many of the early online systems that searched databases containing only bibliographic references. However, a few systems made some provision for document delivery. For example, in 1965, the MICRO system offered a microform collection and a viewer-printer that could be set up next to a terminal. The microform collection had various expanded versions of the records retrieved - abstracts, tables of contents, full source publications. Also in 1965, the LEADER system allowed users to view output at the terminal in the form of document references or of complete textual passages. In 1966, at the Bunker-Ramo Corporation, Canoga Park, California, Van Wente experimented with Bunker-Ramo's version of RECON, which included a capability for online ordering of microform copies of source documents from NASA.
- A choice of simple or advanced user interface accommodates users with varying levels of experience, types of background and preferences for controlling the search process. The capability to select the type of interface was present in 1965 in the MICRO system.
- Access from a terminal in one country to a computer in another enables worldwide sharing and exchange of

information. Global access was first demonstrated in June 1967 when Hal Borko sat at a terminal in Rome and accessed the BOLD system, which was operating in a computer in Los Angeles.

## Web Search Engine Capabilities

Search engines are a common way to find documents on the Web that contain content relevant to a specific word or topic. While hundreds of search engines index the Web, this analysis focuses on the large, general-purpose ones such as InfoSeek, Excite, AltaVista and Lycos. Each of these "supermarket" services indexes about 50 million sites in their databases (although not necessarily the same 50 million). All offer statistically based, relevance-ranked search retrieval, as well as other functionalities. Which of the basic capabilities of online systems are found in Web search engines and how well do Web search engines respond to individual users' needs and preferences?

Boolean operations are common but not universal. Sometimes the logical relationships are automatic or implied; it can be hard for a user to determine when or whether an AND or OR is automatic. Proximity operations and word phrase searching are often but not always available. Natural language query creation is rare.

Fuzzy search is standard. What is rarer to find is the ability to search for an exact match of a query, which may be all that a user wants or needs. Truncation is generally available, sometimes automatically and not under users' control, which may result in unwanted retrieval. Numeric and date ranging, term weighting and field limiting are available, but vary considerably from one search engine to another. Elimination of words on a stop list is common, but it is difficult for users to determine what is on the stop list, which may frustrate a search for certain words or phrases. The automatic incorporation of synonyms into search formulations is common, but not under the user's control to disable when not wanted. Case sensitivity is sometimes available, sometimes not; it is often hard for a user to determine. "More like this" is rare. Citation searching is not found.

Ranking and relevance feedback are available, but based on a variety of criteria unknown to users. Output options are limited in most systems. If zoning is offered, the system usually does not display enough text for the user to make a relevance judgment. Highlighting is occasionally available. The number of records displayed is typically determined by the system and cannot be altered by the user.

Iterative search and canned query are generally not available. Vocabulary browse, concept hierarchies or thesaurus expansion displays are rare. Sometimes, however, expansion is done automatically without a user's knowledge or control.

The Web is an incredible document delivery medium, unthinkable in size and variety just a few years ago. The choice of a novice or experienced user interface is sometimes available. Global access is a given on the World Wide Web.

## Research and Development Needed

In spite of the difficulties in dating "firsts," we can determine that nearly all the basic functions and features were developed for online text retrieval in the 1960s. These "firsts" were developed at many different institutions by many pioneers. Most of the systems of that era were not much more than experimental or laboratory systems with small databases, but the functions and features did actually work and in some cases were applied to significant military, legal or educational applications during that first decade.

In comparing the early developments with the retrieval systems employed on the World Wide Web today, we must be careful not to think of today's search engines as simply more fully evolved versions of the early systems. They do not use all the same functionalities and they incorporate some different retrieval principles. Thus we should expect different performance. For example, the retrieval of thousands of records in response to a query need not be a cause for alarm if a user can see the most relevant ones on the first screen or two. Expert searchers know that the performance of some Web search engines can be improved by changing around the order of entry of terms or by adding or subtracting the number of synonyms and related terms.

Nonetheless, important performance issues remain that deserve more research and development. Few actual users of

Web search engines understand how to manipulate and control a query to maximize the quality of their retrieval. The documentation is limited, sometimes not up-to-date or even nonexistent, and most users are too impatient or unaware to consult it anyway. Thus, despite the vast amount of information that is, in theory at least, accessible via the World Wide Web, most users still retrieve documents that have little or nothing to do with the topic of interest and fail to find the material most pertinent to them.

Furthermore, many search functions and features are done automatically, without a user's knowledge or control. The underlying philosophy of Web search engines seems to be that the system knows best, and users would be well off not to interfere. The Web searcher has given up a great deal of control in exchange for simplicity of use. The result of this philosophy is that, despite its potential, the Web remains difficult for large numbers of users to navigate.

Some of the limitations of Web search engines are being compensated for by the efforts of librarians and other information providers who are creating Web directories to organize sites by broad subject categories or creating metasites of links to related sites or creating publications that list URL addresses of sites important to a certain topic. Other organizations are offering filtering software to eliminate certain types of unwanted retrieval. Others are conducting research on adding metadata to Web records to facilitate intelligent retrieval by man or machine.

In addition to search, browse and miscellaneous other capabilities, the performance of a search system can be assessed in other ways, including completeness of results (recall); finding a needle in a haystack without a mouthful of straw (precision); response time, database coverage, update frequency, output overlap, output options, reliability and friendliness of interface. If we were to revisit the 1960s, we would find that all of these performance characteristics were considered then and rigorous evaluation studies were conducted.

Today's Web designers, developers and evaluators have not yet addressed all these performance issues. In spite of the amazing ability of Web search engines to scan millions of text records of every conceivable type in an instant, users are still frustrated with unsatisfactory results. The apparent ease of use masks the actual difficulty in finding useful information. Designers of search engines of the future must make some critical decisions about how much complexity in document type and content domain they can effectively handle and whether expected advances in bandwidth, language parsing and search algorithms can really address users' needs for useful information. The answer to giving control and power to users might be found in older features such as field specification, synonym incorporation, thesaurus expansion or truncation - or it might not. But remembering the basic functionalities and why they were designed the way they were might reveal just what it is about online text retrieval systems that makes them not only easy to use, but also powerful tools for text retrieval.

---

*Trudi Bellardo Hahn is manager of User Education Services at the University of Maryland, College Park Libraries. She can be reached at McKeldin Library, University of Maryland, College Park, Maryland 20742-7011, or by e-mail at [th90@umail.umd.edu](mailto:th90@umail.umd.edu)*