

# information insider

## What's Next for Text: Retrieval Trends Past, Present, and Push



**Robert J. Boeri and Martin  
Hensel**

**EMedia Professional**, April 1997  
Copyright © Online Inc.

In the new world of Web Time, where every calendar month witnesses demonstrable evolutionary changes in areas of great technological tumult, the lessons of history abound. Full-text retrieval is one such area, and its last half decade is one such epochal time.

About five years ago, purchasing a text retrieval tool meant investing in long pilots of insular systems costing \$100,000 or more. And none of these systems could claim anything resembling "right out of the box" ease and readiness of use. Every system required painstaking customization. Getting documents into the system often required using scanners and error-prone optical character recognition software. And back then, you had to be careful that you didn't let an ASCII document get into a collection of Word or WordPerfect documents or the daily indexing batch process would fail. The Internet was familiar only to defense contractors and academics, and electronic publishing systems such as Corel's Envoy or Adobe Acrobat remained niche tools, if for no other reason than that they were used by so few.

**Once hard to install and arguably harder to use, text retrieval systems have metamorphosed into collaborative, comprehensive, continuous, conceptual, and even cheap searching solutions.**

Accessing a retrieval system across a network was difficult and often very slow, and you needed dedicated IS support to keep the system running. In short, early 1990s-era content-based retrieval used a "pull" model: If you wanted information, you requested it from the system by issuing a query. You then pulled (received) a list of relevant documents in response. Systems were closed, costly, and complex, and they delivered only when explicitly asked.

Flash forward to 1997 and behold the "push" model. Text-retrieval systems have metamorphosed into collaborative, comprehensive, continuous, conceptual, and even cheap searching solutions. Given the sweep of evolutionary change and the transformations it's demanded, it's a wonder any of the original big-three vendors--Verity, Fulcrum, and Excalibur--survived the geotectonic shifts.

Yet survive they did, and those and other text retrieval tools have gone from specialty items to mainstream corporate tools and even desktop commodities. Fulcrum has become the vendor of choice for Microsoft's Exchange. Excalibur and Yahoo! have struck technology agreements. And Verity has aggressively and explicitly pursued general ubiquity for its Topic system. Moreover, these systems have not only grown to accommodate the Web, but they are being profoundly influenced by many Web metaphors--from their user interfaces to their ability to work with and on the World Wide Web.

The way text retrieval's "Big Three" have evolved and continue to evolve provides a fitting window on where text retrieval in general has been and is going.

## **CONCEPTUALIZING THE TEXT: WHAT SEARCHING HAS FOUND**

Conceptual searching has always been a goal of search systems. "Concept-based retrieval," a catch-all term coined by Verity, refers to complex tree-like queries whose branches represent shades of meaning and importance. From the start, these "trees" worked well and easily, but building them demanded expensive handcrafting similar to the knowledge engineering required for expert systems. However, without conceptual searching users are unlikely to find what they are searching for and instead get overwhelmed with thousands of responses or hits.

It was Yahoo!, the Web search service, which first organized searching into conceptual groups like entertainment and sports. Such groupings not only made searching easier but quicker as well, since the engine behind the search had fewer items to process in each request.

Creating the conceptual groups carries its own attendant difficulties, however, and deciding which group or groups a document belongs to for every document in a database can be even trickier. Verity's Search'97 product suite-- using its new 2.0 search engine--can examine documents and automatically categorize them. Find a document you like using the Web Excite search service, and you can ask it to search for more of the same. A similar facility is in the Verity 2.0 engine, making conceptual searching easier.

## **EXTRA-TEXTUAL SEARCHING: A PULL TO PUSH EVOLUTION**

For users who need to search non-textual information in document object collections, a number of capable tools have stepped forward in recent years. Excalibur, whose core search system is based on neural networks, has always offered so-called "fuzzy" searching. Fuzzy searching is very helpful to users searching for words in OCR'd documents which contain spelling errors. Misspelled words could not be found in systems relying on thesauruses to help find similar meanings. However, Excalibur had been slow to apply its technology to visual objects. In December 1996, however, Excalibur announced its Enhanced Visual Search Capabilities for the Internet, which includes an image surfer facility developed by Interpix Software Corporation that will leverage a new partnership with Yahoo! The new capabilities, according to Excalibur, will enable users to search the Internet for photos, diagrams, pictures, cartoons, and other images. Indexing graphical objects requires reducing them to basic attributes so users can pose queries like "Find me something that looks like this."

Another visual metaphor recently proposed comes from Verity. At its recent customer-based Verity Interchange conference, Verity unveiled a graphics-based approach to searching and finding text-based information. Although still in the research phase, conceptual clusters of subject matter can be expressed as two-dimensional maps; clusters appear as cities. The larger the city, the more comprehensive the category; the closer the clusters, the more related they are. Search results are placed on the map, and if they are nearer to some clusters, this gives an immediate sense of their relationship to various concept groupings.

Information collections have always needed to be current, but another Web metaphor has extended the notion of continuous information from a pull to a push model. The Web pioneer which made this model best known is PointCast, a free online news service. Fill out a questionnaire detailing your interests, and each time you log onto the Internet a tailored electronic newspaper is delivered to you. Verity's Search'97 product line similarly allows both retrospective pull searching and real-time profiling and indexing for push queries. And it does this while offering the same document and query analysis tools and user interface. The system acts the same, and achieves the same result, whether you run the query or an active agent does the searching for you. Your agent working for you or fading into the background: this is the ultimate in continuous collaboration.

## WHAT'S NEXT FOR TEXT? COMPREHENSIVENESS AND OTHER CHALLENGES

Since our document information is everywhere, from desktop to the Internet and Intranets, and is found in countless binary forms, the recent retrieval trends matter little if search systems aren't comprehensive. That is, they must be able to find what we want no matter where it is located and what format it is in.

Fortunately, comprehensiveness is another trend that has emerged in 1997. Much corporate information is found in SQL databases, for example, and Oracle Corporation recently announced shipment of its ConText search system. This system allows full-text searching in a relational database, and search results can even be summarized via Structured Query Language (SQL). ConText also allows searching on Intranets and client/server networks, and, of course, searches in ConText can be performed in a variety of western languages and Japanese.

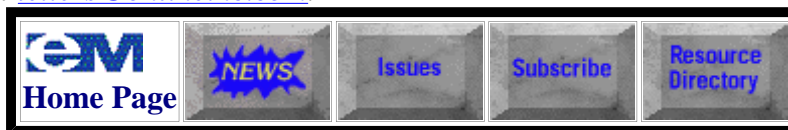
Verity's Search'97 product line achieves similar comprehensiveness via filters and vendor partnerships. Using Inso and Mastersoft viewer-filter technologies, Verity achieves the ability to index, search, and display virtually any contemporary binary format. Partnerships with SoftQuad will provide zone searching in HTML and the generalized native SGML. Lexical analysis in Search'97 is enhanced by partnerships with Xerox and Inso, and Asian partners will provide lexical technologies for managing Asian languages. Lest we forget the freely available search option, you can search Acrobat PDF files in many western and other languages, from the desktop to the Web.

With all the advancements in text retrieval systems to date, challenges remain. Areas where late-breaking improvements are showing up or future changes are needed include speed of performance, which Verity, for example, has improved by hiring Web Search programmers to optimize code. Other challenges include indexing increasingly massive document collections, ease of constructing queries or instructing search agents, and the ability to sift through massive document infobases and deliver the best possible answers to our questions. In short, the search for text retrieval goes on.

---

**Robert J. Boeri** and **Martin Hensel** are columnists for *Information Insider*. Boeri is Advanced Systems Specialist in the Information Services Division of Factory Mutual Engineering of Norwood, Massachusetts. Hensel is founder of Martin Hensel Corporation, a Newton, Massachusetts-based consulting firm that builds SGML-based editorial and production systems for publishers, corporations, interactive services, and compositors.

Comments? Email us at [letters@onlineinc.com](mailto:letters@onlineinc.com).



Copyright © 1997, [Online Inc.](http://www.onlineinc.com) All rights reserved.  
[info@onlineinc.com](mailto:info@onlineinc.com)

[This site created for best results under Netscape.]