enhanced PDF collections on the web

chapter eight

The concurrent releases of Adobe Acrobat 3 (optimized files) and Verity SearchPDF combined to create a watershed of digital document distribution technology. Richly enhanced collections can now be served up on the Web with a luxurious set of search features.

The non-intuitive market strategy of the Internet has led to many free search engines. The Excite engine from Architext is a prime example of powerful software technology given away free. It doesn't seem to make sense, but consider the Acrobat Reader, the Netscape Navigator, the Microsoft Internet offerings; they have all been free.

And the most remarkable feature of the free products from the big commercial vendors is simply the fact that they are being offered by commercial, most definitely for-profit corporations. But the Web itself was born of free software, the NCSA and Cern servers, and the Lynx and Mosaic browsers.

The single biggest problem with free software is that users are required to download and install new versions, and to a large extent support themselves. (Adobe experiences 20,000 to 30,000 downloads of the Acrobat Reader per day!) A free, fuzzy search engine may offer little value to users of a collection because a commonly accepted organization is lacking. If users don't understand how the data is structured, their searches are just shots in the dark.

From a Web operator's point of view, the most common failure is in hardware and software support. But from a user's point of view, the major concerns are rapid response and useful results.

A person seeking information is probably fired up, motivated, and therefore impatient. A search that takes a minute or more to process is simply insufferable to many online users.

A carefully constructed Catalog-indexed collection will offer excellent performance from the user's point of view. The motivated user will always enjoy the best that any information collection has to offer. Search aids and descriptions of the indexes will help more users become expert users. Searches that specify Document Info fields such as Author, Title, Subject, Keywords and System Info Fields (Creator and Producer, Date Created and Date Modified) will be very fast and productive.

The search capabilities available in Acrobat 3 to query the entire database offer a uniquely disciplined and organized Web resource. Directed to the needs of a specific audience, there are few if any comparable packages for creating a digital document database at off-the-shelf prices.

Full Power Searching On The Web

Version 1.0 of Verity's SearchPDF for Web servers offers three standard user interface screens for entering queries. They are Simple Search, Standard Search and Power Search, and they appear to be dead-simple easy.

Through the Simple Query entry window, the extremely sophisticated text-searching technology is fully available. All of the logic operators are available to be entered into the Query field, including the Evidence, Proximity, Relational and Concept Operators. This means that this free software can serve the needs of a supremely demanding audience for information retrieval. The only caveat is that the users must learn how to use the system to appreciate its advantages.

tip

The two measures of information retrieval are precision and recall. If a query recalls every document in the database, the user will never find what he is looking for. On the other hand, if the user can't specifically identify the query term exactly as it appears in the collection, relevant documents are lost.

However, in large collections, the first problem is usually most difficult, where too many irrelevant documents match the search criteria, leading to a failure of precision. Just a few operators can enable a user to be tremendously more productive, more precise, in searching text collections, by narrowing down the retrieved documents to only those most likely to be relevant.

Simple Search (top left)

If additional collections of catalog-indexed collections were available, they would be listed under the Select Collections in a check box list.

Standard Search (center)

Standard Search offers enhanced options over Simple Search, allowing for more complex queries.

Power Search (top right)

All of the Verity topic query operators and syntax may be exercised through Power Search, and the user may select the maximum number of documents retrieved.

The Basic Boolean Operators

Since the dawn of full text retrieval, these functions have been the building blocks of information retrieval from unstructured databases. Because these concepts go so deep to the core of free text searching, many users expect these functions to work on every database. Through many variations on a theme, information seekers naturally gain proficiency in these Simple Search techniques.

Brackets should be used to join query terms and operators into a single search argument. Boolean operators are:

And	**< Near >**
Or	**< Phrase >**
Not	**,**

example

((Kelly Near Johnson) Or (skunkworks))

Near (Phrase “faster than a speeding bullet”)

Or (blackbird, aurora, stealth, SR-71, U-2, F-117)

Not (thrush, borealis, ninja)

Prefix And Infix Notation

”Words that use any operator except evidence operators (Soundex, Stem, Wildcard and Word) can be defined in prefix notation or infix notation.

Prefix notation is a format that specifies the operator comes before the words to be used with that operator. The following example means: Look for documents that contain a and b.

And (a,b)

When prefix notation is used, precedence is explicit within the expression. The following example means: Look for documents that contain b and c first, then look for documents that contain a.

Or (a, **And** (b,c))

Infix notation is a format that specifies the operator be between each element within the expression. The following example means: Look for documents that contain a and b or documents that contain c.

a **And** b **Or** c

The logic of infix notation is that each operator appears between each element, which means that the section “a and b” (where “a” and “b” are the elements and “**And**” is the operator) is executed before the next section in the statement, which contains another operator and element (“**Or** c”).

When infix notation is used, precedence is implicit within the expression; for example, the **And** operator takes precedence over the **Or** operator.”

Quoted from the Verity SearchPDF online documentation.

Evidence Operators

The evidence operators provide term expansion, which takes advantage of the Word Stemming, Wildcard and Soundex options of Acrobat Catalog. The author or publisher of the index must include these features during index building, or they will not be available for SearchPDF queries. Evidence operators are:

Word	**Wildcard**
Stem	**Soundex**

Proximity Operators

Proximity operators allow the user to specify that the query terms must appear within a certain distance of one another to constitute a highly ranked hit document. For example, the Near/N operator can specify that terms appear within a specific distance of one another in the text. Proximity operators are:

In	**Paragraph**
Phrase	**Near**
Sentence	**Near/N**

example

A query delimited with the **Near/N** operator can find related terms in multiple orders:

The query "Kelly **Near/2** Johnson" finds Kelly Johnson; and Johnson, Kelly; and "Kelly" (Clarence) Johnson.

The query "Kelly **Near/2** Johnson" will not find the following relevant article because the Query Terms are too far apart:

> Clarence *Johnson* is the legendary designer who visualized the most advanced aircraft in history in his head, and then he managed a cadre of engineers in the Lockheed Skunkworks to build them. In the case of some of his inventions, he had to build the factories that could build the planes. The peerless SR-71 Blackbird had to be fabricated in titanium to survive the performance regime that "*Kelly*" envisioned. Titanium had never been worked before, and all new tools and procedures had to be created to handle this unusually strong element.

The previous paragraph contains interesting material that is very pertinent to the user's intended query, but it would never be retrieved with the narrow proximity operator of "Kelly Near/2 Johnson," which requires the two terms be within two words of one another. Some search engines offer proximity operator of within sentence, within paragraph, within "X" words and so on. A **Within Paragraph** operator would find this reference.

Relational Operators

These operators are designed to take advantage of the preparation of the files in the collection by providing the means to selectively use the information in the Author, Title, Subject and Keyword fields. Relational operators are:

Contains	**Starts**
Matches	**Ends**
Substring	

tip

Don't forget the extra operators:

? Single character wildcard

*** String wildcard**

' Single quotes initiates word stemming

" Double quotes finds only exact matches

Concept Operators

Concept operators are the glue that congeal many query terms into a model of an idea, or a "topic," in Verity-speak. Very precise arguments, or query statements, that have been built with the evidence, proximity and relational operators can be synthetically combined into one large search "equation." Concept operators are:

And

Or

Accrue

tip

Verity SearchPDF is just the tip of a family of information-retrieval tools called topic (with a lowercase t). SearchPDF is a sophisticated and very effective teaser for the rest of the software tools. There are a couple of key limitations or enhancements available in the free SearchPDF package. While all of Verity's extensive search and retrieval functions are available, they only work on Acrobat Catalog-indexed collections of PDF files. Verity's engine can actually index and search hundreds of formats, and users who enjoy searching PDF documents may feel a need to have the same search power over other formats. That's where the rest of the topic family comes in, offering the ability to search virtually all digital documents.

Another key capability reserved from free offerings like SearchPDF is the ability to reuse queries. In fact, the product name topic refers to the ability to precisely define a concept, or topic, that embraces a complex question. These topic queries can be stored and executed on schedule or on demand. Expert queries can be adapted and reused indefinitely on dynamically changing data.

A topic persistent query can constantly interrogate data sources, such as stock tickers, wire services and many others, and then generate timely reports of new information that appears relevant to the information seeker's interests. This is called the topicAGENT.

Intelligent agents, persistent searches and other automated information-gathering techniques offer viable means to surf the overwhelming deluge of information coming online.

Modifiers

Modifiers allow the user to put a particular spin on each of the search terms. These modifiers give the savvy user the tools to create queries that can deliver traditional reports of results, including the "order" modifier. Modifiers are:

Case For word & wildcard searches

Many Density of term vs. length of document

Not Exclude "thrush" when searching for "blackbird"

Order Specific order of terms in sentences, paragraphs or proximity

The Verity topic products provide for storage, reuse and flexible deployment of these pre-defined topic queries. Very advanced information-retrieval techniques are available to tailor Web topicAGENT software robots to roam the Web or specific databases and bring back the most desirable and valuable information.

tip

Many times, an initial search on a database will provide clues on which terms to exercise the NOT operator. For example, if you were searching a news database for the "blackbird" spy planes, you would like to avoid mention of the English blackbird, which is actually a common European "thrush." By excluding this term, you would avoid mentions of blackbirds in nature and birdwatching articles.

example

Density refers to the number of hit terms as a fraction of the entire document (hit terms/total terms). Logically, if the hit terms comprise the bulk of the document, it is likely that the hit terms are the most relevant subject of the hit document. This feature is specifically designed to overcome frequency errors. For example, a 400-page manual will probably contain a greater number of hits on a certain term, but the information will be more diffused than a 400-word post that contains a lesser number, or frequency, of hits.

For example, to search for the best solution for both paper and online publication, you might query for:

"Best fonts and point sizes for both paper and online publishing"

An authoritative 200-word listserv posting on PDF-L will mention the terms in logical order with very little extraneous data. Therefore, the ratio of search terms to all terms in the individual posted document will be very high. The answer to the original question is likely to be given in this type of hit.

On the other hand, a long document, or even a list or index, may contain many more instances (frequency) of the hit terms, but at a much lower density. The user is more likely to find information in the shorter, "denser" document. Of course, if not, the longer documents offer another option.

tip

The Many modifier directly affects relevance ranking. If the new user were to learn only one "extra" command, it should be the Many operator. In most texts, this will lend extra importance or weight to a particular term.

example

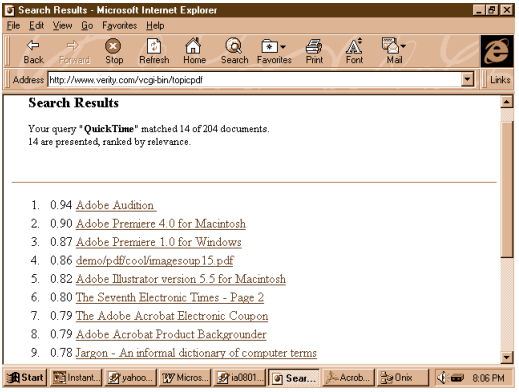
"< MANY > supersonic AND stealth" would retrieve all information about supersonic and stealth aircraft, and would present supersonic stealth aircraft at the top of the heap.

Check out SearchPDF for yourself on the Web. Excellent documentation is available on all search features:

http://www.verity.com/demo/index.html

Simple Search

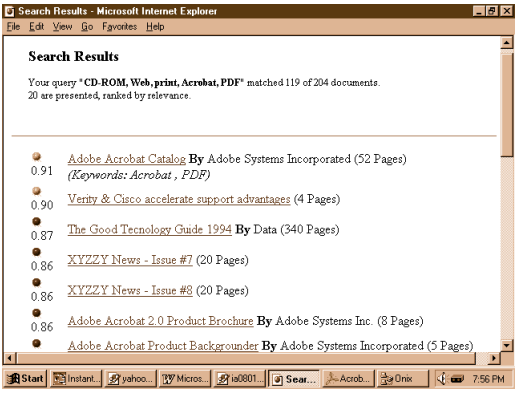
The only options presented here are the ability to select collections for searching, and the Query window. The zen of this interface is that the entire organization of the database is accessible through that simple view. An experienced user can enter the most highly structured queries into the elementary screen.



A list of hypertext links is returned as the hit list of a Simple Search, with documents ranked numerically by relevancy rating.

tip

For Simple Search, Standard Search and Power Search screens, see the first section of this chapter.



The Standard Search hit list returns document information such as Author (By), File Size and Keywords.

Standard Search

The Standard Search screen offers two pushbutton qualifiers for the terms entered into the Query field. The default choice, “Words or phrases separated by commas,” is the approach taken in the Simple Search above. The other choice, “Free-form text,” offers the user an unstructured interface that interprets the information entered into the Query field.

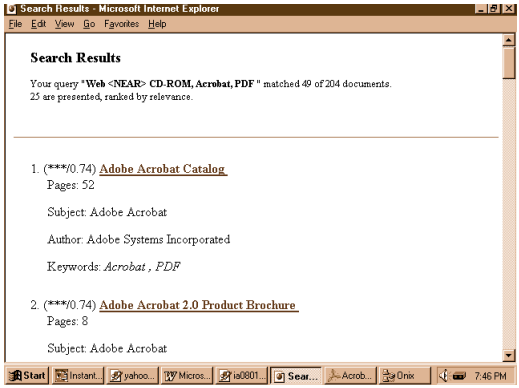
Power Search

Power Search offers control of the number of documents retrieved by any query. There is no need to clutter the user's mind, the network's bandwidth or the server's CPU with excessive file transfers. This option provides economies on every step of the info-transfer process by simply limiting the number of documents referenced and handled for retrieval.

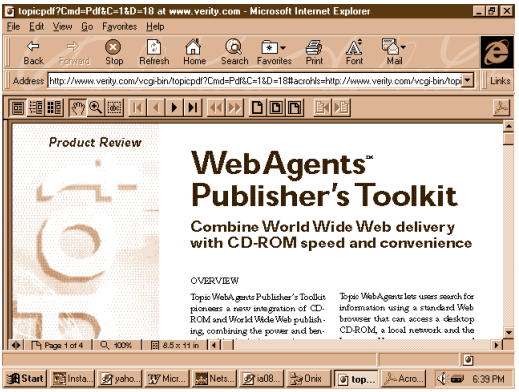
tip

Remember, while you are searching for words or ideas, you are actually retrieving documents. The best query retrieves the most relevant document by prioritizing the overall concept — through careful use of search operators.

Every query should consider relevancy ranking because that determines the order in which the retrieved documents will be presented. Once again, it's the question of precision and recall. By weighting queries toward precision, the desired information floats to the top for easy retrieval.



Power Search retrieves document info that could be exploited by a careful author or publisher who knows the needs of his users.



When a PDF document is selected from the HTML link in a results list, the Acrobat Plug-in automatically views it within Netscape Navigator and Microsoft Internet Explorer, as shown here.

Web Servers and SearchPDF

Verity's strategy for this great freeware is consistent with the Internet model, so SearchPDF is designed to run on the most common and popular server software platforms. Version 1.0 runs on NCSA, O'Reilly's WebSite, Netscape's Commerce Server and Microsoft's Internet Information Server.

Search Results In HTML

The Search Results page includes a relevancy ranked list of documents that match the search criteria. A numeric relevancy grade is displayed, as well as the contents of the Keyword field from Document Info.

This simple combination offers the best of both worlds: Relevancy is computed in the software as it analyzes a document, while the Keywords let the author or publisher specify relevant interests in the Keyword field for each document.

The PDF documents themselves are available via hyperlinks in the hit list. It is very considerate of the Webmaster to note the number of pages in the file. You might have time to grab a 4-page PDF, but you'd have to carefully schedule downloading a 400-page file.

There are actually six wildcards:

?, *, [ ], { }, ^, and -

All of these functions are documented online and, once learned and found useful, will be employed by future users as easily as we employ arithmetic now.

The Power Search Results page of SearchPDF returns the hit list with the following helpful fields:

Relevance Rank, Hyperlink Document Title, Author, # of Pages, and Keywords.

Experience teaches users to instantly recognize patterns and information in the hit list, and this presentation provides for efficient browsing of the hits.

tip

Help the User by Explaining the Organization in the Collection

For example, a manufacturer offering digital documentation might provide online users with these helpful Hints on the home page:

Reminder - Always use the "Contains" operator in Field Searches for widest possible retrieval. That is, use Parens to include Field Names with Contents, as in (Subject ~ WidgetX3). These searches will retrieve all documents with your search term in a particular field.

Reminder - We use each Info Field as follows:

Title - Document Name, including such useful search terms as Install Manual, User Manual, Online Help, etc.

Subject - Product Name, Product Number

Author - Product Manager, Corporate Contact

Keywords - Acronyms, Nicknames, common references, etc.

Multi-Platform Access

Operating systems mirror the primacy of platforms on the Internet and in corporate America. Windows NT is the fast-growing "new install" by far, while a significant part of the installed Client/Server environment runs on Sun Solaris, HP-UX and IBM AIX. SearchPDF runs on all of these platforms and is compliant with any Web server that supports a Common Gateway Interface (CGI).

CGI allows otherwise static and one-way HTML pages to provide dynamic interactivity. The basic Web browser is designed to follow links and display pages. CGI scripts are the links between these hypertext pages and all other complex online processes. They are programs that run on a Web server, and they are initiated by a Web-linked user. Any anonymous surfer on the global Web can run a program on a remotely linked server through these commands. This facility is universally exploited on the Web to offer enhanced functionality.

Today, Java and ActiveX offer the next generation of dynamic functionality originally provided by CGI scripts.

Acrobat 3 provides extensive, fluid access to the PDF content online, including the display of hit terms in the retrieved documents.

Verity's SearchPDF is specifically designed to work with indexes built by Acrobat Catalog. SearchPDF does not offer traditional Verity text-retrieval capability on other collections, such as all the word-processing files on your hard drive.

Sony Electronic Publishing Services Tackles Large Jobs

case study

Bob Marsh, of Sony Electronic Publishing Services, a division of Sony Disk Manufacturing, has mastered techniques for handling one of the largest CD-ROM publishing applications to date. In the spring of 1995, SEPS won the contract from the Institute of Electrical and Electronics Engineers to convert over one million scanned images to Acrobat format and publish the entire collection on CD-ROM.

The collection of .tif files going back to 1986 included conference proceedings and colloquia, standards and journals, and was housed in a proprietary viewing system and directory structure. The project included re-mastering 220 CDs.

"We had three primary goals in the project," Bob explains. "First, we wanted to upgrade to a non-proprietary environment, and we chose Windows. Secondly, we decided to use the Verity Search engine. Finally, we wanted to move from scanned input to electronic source material."

"We decided to use Acrobat at the outset. We now use Acrobat on everything we do for several reasons. It's a cross-platform format that is efficient for the widest range of users. We can integrate scanned and electronic source material in a way that is totally seamless to the end user."

"The converted images are searched through the meta-data provided by IEEE, which is tagged ASCII in their INSPEC database, which is widely used by technical libraries. We created our index from the ASCII and attached the images."

"The project included three systems we developed for the conversion, production process and retrieval tools. We provide retrieval through a VB wrapper on top of the full Verity engine and Acrobat serves as the embedded Viewer. The path to the articles is through a hierarchical browser, which works through a tab dialog system."

"We use Verity to search the INSPEC data which indexes the database of images. The index requires two CD-ROMs, and the collection of a half million articles is published on 240 CD-ROMs."

"For simplicity's sake, we decided to deal with only scanned images, and provide the means for users to take advantage of the meta-data professionally compiled by IEEE to search and retrieve documents. Going forward with monthly updates, we now take electronic source in PDF directly into the system, eliminating the need of scanning and image conversion. The collection appears as a seamless whole to the end user."

— Thanks to Bob Marsh
of Sony Electronic Publishing Services

Visit the web site at: http://www.seps.com

Summary

One author using Acrobat 3 can convert many forms of paper and electronic documents into PDF content. Very large collections of files can be processed by Catalog so that a single, comprehensive index offers instant access to information content.

Verity's free SearchPDF allows for up to four indexes, providing the potential for information-retrieval capability over literally millions of pages.

A Webmaster should publish a directory of the conventions used in the collections so that even newbie users could efficiently peruse the information.

The full Verity text-search functionality conveys dynamic research functionality upon any large collection of PDF documents on the Web.

Offering such capability on either the Web or an Intranet should not be taken lightly. As always, user satisfaction must be of paramount importance. It is always a good idea to overbuild every aspect of a Full Text Retrieval database, including the CPU, disk drives, communication interfaces and available RAM. Remember that the cost of the server is spread over all of the users of the system.


CALIFORNIA
US
66