

Document Management System Search Techniques: The More, The Merrier

Introduction

More and more systems are adding more and more search features; soon all of the systems will have all of the features. What are these features, and is this gathering multitude a good thing?

Yes. Having a full suite of search techniques is like extending our personal way of doing business to the way we work with the computer. When we start a search ourselves, we use everything we have. Document management systems, that provide a complete search suite, allow us to extend our personal 'full-court-press' to the use of computers.

Types of Searches

A search with a key, such as "Please deliver record storage carton (box) 8192" does not seem like a search at all, because it is too easy. "Please find me everything we have done like this." is indefinite, uncertain, but we can respond. "Change 'tired-old-marketing-term' to 'hot-new-marketing-term', everywhere." is specific and is easily done with a computer. Then there is the 'smoking gun', the goal of litigation support searches. The 'smoking gun' is the easily presented documentary evidence that provides absolute proof of an illegal act.

In short, there is nothing that cannot be requested, or expected.

Here are some search techniques, from old-as-the-hills to maybe-someone-is-doing-it-but-we-are-not-doing-it. Here are some search techniques from the time honored past to the promising future.

Sequential Search / Date Scanned

We search for words in a dictionary sequentially, after we get to about the right place. We search sequentially for folders in boxes and for documents in folders. Sequential searching reduces the 'up-front' cost of indexing and places the burden on the searcher. There is little wonder that it is the most common search technique.

In books, a craft called book design allows us to flip through the pages of a book at up to 60 per second. The pages of books are actually designed to be viewed at 60 pages per second. Computer systems will reach this speed when the DVD (commonly Digital Video Disk) is fully integrated into PC design. When PCs can display pages at 60 per second, document imaging will be able to offer some of the most important searchability benefits of paper. Chief among these benefits is the shifting of indexing costs to the searcher through high speed sequential searching.

Document imaging's equivalent to the sequential search is the date-scanned search. Very often, when a document cannot be found, its scan date is

estimated. The most likely time period during which the document might have been scanned into the system is established by reviewing the organization's procedures. Then a sequential search is mounted in which all documents scanned in during the most likely input period are searched sequentially.

Nested Folders / Aliased Folders

Windows 95 brought the folders known to Mac users (and nostalgic Xerox Alto and Star users) to the attention of the other 90 percent of PC users. These folders, like folders in a file cabinet, arrange groups of documents in a sequential filing arrangement. Like the folders in a drawer, drawers in a file cabinet, file cabinets in a file room, and file rooms in a corporate division; the Windows 95 folders have a nested, hierarchical arrangement.

To accommodate the common problem of wanting to file a folder in two places, Windows 95 has aliasing, called short-cuts, which creates a ghost copy of a folder in each place it could be filed. These ghost or alias copies work well until the actual folder is removed, then the aliases are orphaned and problems ensue.

Windows supports ordered sequential folders arrangements. Folders can be arranged by name, by date, by type, or by size. The arrangement can be adjusted at the click of a button, unlike paper folders which must be rearranged manually.

Database Entry / Record Key / Unique Identifier

When records have been indexed and maintained well, searching for a specific key produces the desired results. Searching an employee database by social security number provides a guarantee of finding the one matching employee, or the assurance that the person with that number is not employed by the organization.

In our zeal to make the best system, we often seek the high quality of a unique identifier based indexing system without the budget to pay for the up-front costs of its implementation. This results in increased expenditures and search results that have a lower than expected quality.

Key Words / Subjects

Key words mean up-front work. Someone must pick the keywords for a document. Synonyms (such as 'box', 'carton', and 'container') must then be tolerated, or the system must be designed to eliminate synonyms. To eliminate synonyms, there can be a requirement to select keywords from an approved list, known as a controlled vocabulary. The controlled vocabulary must be kept up to date, and disagreements must be resolved.

There may be hierarchy of terms which may evolve into a subject listing. For example: dog

and cat are animals. Grouping related terms for searching is also called a concept or thesaurus search. The hierarchy adds another level to maintenance and training requirements and costs.

Budget cuts are particularly hard on subject / key word systems. Documents indexed after a budget cut, which eliminates the persons assigning the key words, do not have key words assigned. These documents disappear from the perspective of persons using key word searches. At this point, all searchers must understand that some of the documents do not have assigned key words. This increases training costs during a time of shrinking budgets. Very quickly, it becomes easier to disable the key word search function rather than dealing with its growing complexity.

Conversely, subject indexing provides the best control over documents. There is nothing that compares to knowing what you have in your documents. As organizations become more and more information based, the stored documents become the primary assets of the organization. Husbanding these resources will become a recognized fiduciary responsibility within organizations. This should increase the use of subject indexing.

Full Text / Fuzzy Search

Full text searching has become well known now that word processors have 'find' and 'replace' functions. Full text search is popular because it can be fully automated and shifts costs to the searcher, eliminating most up-front indexing costs.

Because words have various forms, a wild card feature has been added to some full text search systems. With a wild card Sm*th* will match Smythe and Smith. To handle even more variation, fuzzy searching will match misspellings and a wide variety of word forms and idioms. 'Xmas' is a fuzzy version of 'Christmas'.

To further refine searches, several variations on word counts have been added as options. Examples are: the number of times a word appears, the appearance of more than one search term in a document, the order of the search terms, the distance between search terms, and the appearance of search terms in 'important' places in a document, such as the beginning or abstract.

Full text indexing has the advantage that it is not affected by a prevailing point of view at the time the document indexing was done. For example: oil companies have traditionally indexed their documents for the purpose of finding oil. When pollution control became important, these indexes were of no value. Full text indexing, on the other hand, provides search hits on any subject contained in indexed documents. (In computer based searches, identified references have become known as 'hits'.)

Full text searching is also used in COLD (Computer Output to Laser Disk) and COOL (Computer Output On-Line). These document

files are simply print images of computer reports, but, full text indexing frequently provides faster access than the original organization of the report.

The best example of eliminating up-front document processing and indexing costs is full text indexing of raster scanned documents. (Raster scanning is the process used in fax machines to convert documents to the black or white dots called pixels that are transmitted as ones and zeros over phone lines.)

The raster is converted to text using OCR (Optical Character Recognition). The resulting text is never perfect, but produces many useable hits in full text searching. The key to assessing the value of full text indexing of raster scanned documents is to compare the results to the alternative -- a manual sequential search of the paper documents. This assessment easily explains the huge popularity of full text indexing of bulk-scanned documents.

Card Catalog / Finding Aid

Existing card catalogs can be scanned in and full text indexed. This makes past expenditures in subject indexing available at very little cost. Many records series are indexed on 3 by 5 inch cards. These cards are called finding aids when the corresponding records are accessioned into archives. These finding aids can be scanned and full text indexed as well.

Document Structure / SGML

In searches, search terms can be given more weight if they appear in 'important' parts of documents. Examples of 'important' parts of documents are: titles, abstracts, introductions, summaries, and illustration titles. These document parts are identified in SGML (Structured Generalized Mark-up Language), the language of structure for manuals and procedure books. HTML (HyperText Mark-up Language), used on the Internet, is a subset of SGML. XML (Extensible Mark-up Language) is an effort to expand HTML toward the full capability of SGML. All Microsoft Office 2000 documents will be stored in XML format.

In SGML, the structure of a document is identified first, and the text is then flowed into the structure. This is a common distinction in desktop publishing programs such as Adobe FrameMaker where a document is designed and then the text is entered from a popular word processor. SGML based products are the answer to the question: "How can I avoid having to re-do my formatting over and over again?" And "How can I get everyone to use the same format?" As Microsoft Word keeps adding features, the internal Word document structure will have to be SGML-based to handle the complexity. The latest step on this road is for Office 2000 to store documents in XML format.

Hyperlinks

Hyperlinks are the terms or phrases underlined in blue on Internet web pages. Clicking on a

hyperlink jumps to the referenced page anywhere in the world. Each month hyperlinks require less and less description as more and more people start to use them. Word processors and document management systems are quickly adding the capability to added hyperlinks to new and existing documents. Because all Office 2000 documents will be in XML, any Office 2000 document can be saved to an Internet web site for search and retrieval access by the entire world via the Internet.

In the future hyperlinks will be annotated and ranked by quality. The number of people following a given hyperlink will be counted. Pages will be ranked by the number of quality hyperlinks going to the pages. All this 'meta-data' (information about information) will be useable in searches. Fortunately, one of the design criteria for hypertext links is that they should not get in the way if you are not interested in them. Current examples of hyperlink meta-data are the citation indexes and bibliographies published in book and database form.

Workflow links and emailed links are among the many special cases of hyperlinks that will continue to appear in the future. Internet agents and popularity charts will add to the multidimensional hierarchy of meta-data that will enhance existing search options.

Image Matching / Physiological ID

The same fuzzy searching used for text also can work with images. For example, what does this logo look like? (Please insert a logo of your choice.) (This is a note to the reader, not to the editor.) When you look at a logo, you compare it to every image you have ever seen, simultaneously, instantly.

Your comparison is part of the definition of your looking, of your seeing. It takes no extra effort, and it is not sequential. It just is. This is the working of a neural network, which is what we all use. This neural network can be modeled in a computer, and is the basis for some commercial image and text search products.

Applying this technology to fingerprints, eyeprints, voiceprints, etc., extends searching to individuals.

GIS and CAD

GIS (Geographic Information Systems) are the modern day maps that show every aspect of our world on a multi-layered three dimensional grid. GIS systems give us the ability to search for things that are 'close' to other things. Adding CAD (Computer Aided Design), we can search for things that are physically like other things in three dimensions, by composition, or manufacturing process. We can find the plans to our house, and see if they conform to the current zoning regulations. We can pick a spot, parcel, or city and go forward and backward in time and watch the process of construction and city growth or decline. We can search for a house of a given size, that was built during a given period, near a given area.

SMPTE Time Code and GPS

The SMPTE (Society of Motion Picture and Television Engineers) is responsible for the time code that specifies when every frame of every movie and every video was shot.

With the GPS (Global Positioning Satellite System) all hand-held cameras will be able to continuously identify their location. Every picture will have the date, time, location, camera angle, and focus included with the image. Voice annotation will include narration with the photos.

Many of these features are already being used to create the digital orthophotographs that are used to drape the image of a city over the three dimensional DTM (Digital Terrain Model) of the land to provide a visual orientation for searchers. As the price of computer chips drops, even paper disposable cameras will have these capabilities, just as paper greeting cards now have chips that play music.

Using GPS and inertial navigation, a few photographs of a building will result in a three dimensional model that will show the variances between the as-built structure and the approved CAD design of the structure. Appropriate modifications can then be made, either to the structure or to the approval through an amendment.

Log of Reading History

Computers can easily keep track of all documents viewed (and read) by a searcher. This will greatly simplify searches. For example, there are many documents that reference trees. There are far fewer documents that reference trees that an individual searcher read last summer.

In the same way that we find it easy to see where someone is looking, a computer can use a video camera to watch what a searcher actually reads, so your log of reading history can actually contain which words you read, in what order, and how long you looked each word. This is good for searching, but bears review for reasons of personal privacy.

Combination of Techniques

In any search, the searcher is always building a plan of attack, forming a mental picture of the information available and of the information being sought. Any new piece of information is added to all previously found information and forms the basis for the next step, the next search in a infinitely extensible series of searches that makes up a given search. Each search technique turns up different aspects of the information sought, providing new avenues along which to search.

In a successful search, the focus is always on the person doing the search and their thought process. The person forms an opinion and makes pronouncements based on the overall results of all searches done. The person provides answers; searches and search systems do not.

Saved Searches

Searchers do not just use current searches in their quest. Searchers use what they remember from every previous search they have ever done. By saving previous searches, search systems automate and extend this use of previous searches.

These saved searches can be made available to other searchers. These saved searches form another facet of the search repertoire that will soon be available to everyone, using every search system.

Summary

Many search techniques have been invented. Gradually, all of them are being made available in every document management system. The best technique is to evaluate all techniques for their applicability to the search at hand and then to use all applicable techniques in combination.

As computers become faster, and search techniques are adapted to the capabilities of current computer configurations, search times will shorten to become effectively instantaneous. The quality of the search will then depend on the searcher's knowledge of the search problem, knowledge of the broad range of search techniques, and creativity in applying the techniques to the problem.

Sidebar One

Binary Tales

Many of the best things in life are binary. For books, we have two sides to every sheet, giving each leaf a recto (front) and a verso (back) page. When we fold a sheet of paper we get a folio signature of two leaves each of which has two sides producing four pages. Folding the sheet again we get a quarto signature which, when the bolts (uncut edges) are trimmed on three sides (head, and tail or foot, and fore-edge), produces four leaves with eight pages. Folding the sheet again we get an octavo signature of 16 pages. Folding again we get a sextodecimo or 16mo (time honored, but not elegant, trade term) of 32 pages. Folding again, a 32mo has 64 pages. Folding again, a 64mo has 128 pages. And now we come up against the curious fact that a piece of paper cannot be folded more than 7 times. To go where no one had gone before, we can gather the signatures into a book and bind them. (Looking down the spine of a book, one can see the remains of the uncut edge of the signatures in the binding.) Gathering two 64mo signatures, one now has eight doublings that produce 256 pages, the same count as the 256 characters in 8 bit ASCII ((ANSI (American National Standards Institute) Standard Code for Information Interchange).

In ASCII, one bit distinguishes each of two characters in a group of two characters, one character is identified as a 'one' and the other character is identified as a 'zero'. Two bits distinguish four characters in two groups of two

characters. Three bits distinguish 8 characters in two super groups of two groups of two characters (8 characters all together). This progression can be continued to 8 levels of groups distinguishing 256 characters. The binary nature of printed books shows that Gutenberg preceded the Buccaneers, and their pieces of eight, in the binary tales.

Sidebar Two

Marketing a Fuzzy Concept

Fuzzy searching was first marketed by vendors of neural network based systems. These network systems used a computer model of a neural network, such as the neural network that a frog uses to see and catch a fly, to locate search terms.

Other search vendors studied language structure to locate search terms.

When it became apparent that the search benefits of neural networks were selling, vendors of other search techniques gave new marching orders to their marketing departments. The marketing departments were ordered to 'do something'.

The astute marketing departments immediately set out to learn how to spell 'fuzzy' and put the word 'fuzzy' in their brochure. Simultaneously, the marketing departments sent a request to their engineering departments to find out what 'fuzzy' was and how it related to the search product they were currently marketing.

The marketing departments then identified the search technique in their current product that was most like fuzzy searching. In many cases this was the wild-card or "*" search. The identified similar technique was then designated as the 'fuzzy' product offering in the sales training material so that salespersons could respond to prospects who requested an in-depth presentation of the 'fuzzy' product. This designation completed the development process for the first release of the 'fuzzy' search product and assured that the truth-in-advertising requirements were met. The marketing departments then requested a follow-on technical solution from their engineering departments. If funded, these follow-ons could be used in subsequent product releases.

Updates and More Detailed Descriptions

When using the information in this article, please check the website www.ArchiveBuilders.com for updates. The version number of this article is just before the page number below. The website also has articles that provide more details on some of the terms and concepts in this article.

Comments

Please let us know how you like this paper, or if you had any questions. What would you like to see in the future? For more, and the most recent version of this article, please visit our web site at www.ArchiveBuilders.com.

Please send your comments via email to SteveGilheany@ArchiveBuilders.com. Tel: +1 310-937-

7000 Fax: +1 310-937-7001. Also, please let us know where you saw this article.

Acknowledgements

Reprinted from *Archive Planning*, Volume 2, number 6, 1998, Archive Builders' analysis newsletter for document management.

See www.ArchiveBuilders.com.

All trademarks are the property of their respective holders.

Note to Editors

Paper 22005v010

We will continue to update these articles as we get comments. Please contact us for the most current version before you publish. Also, please request permission to publish the article. Permission will be given freely for most purposes.

Steve Gilheany
Archive Builders
1209 Manhattan Ave., PMB C-14
Manhattan Beach, CA 90266
Tel: +1 310-937-7000 Fax: +1 310-937-7001
SteveGilheany@ArchiveBuilders.com

Bio

Steve Gilheany, BA in Computer Science, MBA, MLS Specialization in Information Science, CDIA (Certified Document Imaging System Architect), AIIM Master, and AIIM Laureate, of Information Technologies, CRM (Certified Records Manager, ARMA) has seventeen years experience in document imaging and is a Sr. Systems Engineer at Archive Builders.

Author

Steve Gilheany is a Sr. Systems Engineer at Archive Builders. He has worked in digital document management and document imaging for seventeen years.

His experience in the application of document management and document imaging in industry includes: aerospace, banking, manufacturing, natural resources, petroleum refining, transportation, energy, federal, state, and local government, civil engineering, utilities, entertainment, commercial records centers, archives, non-profit development, education, and administrative, engineering, production, legal, and medical records management. At the same time, he has worked in product management for hypertext, for windows based user interface systems, for computer displays, for engineering drawing, letter size, microform, and color scanning, and for xerographic, photographic, newspaper, engineering drawing, and color printing.

In addition, he has nine years of experience in data center operations and database and computer communications systems design, programming, testing, and software configuration management. He has an MLS Specialization in Information Science and an MBA with a concentration in Computer and Information Systems from UCLA, a California Adult Education teaching credential, and a BA in Computer Science from the University of Wisconsin at Madison. His industry certifications include: the CDIA (Certified Document Imaging System Architect) and the AIIM Master, and AIIM Laureate, of Information Technologies (from AIIM International, the Association of Information and Image Management, www.AIIM.org), and the CRM (Certified Records Manager) (from the ICRM, the Institute of Certified Records Managers, an affiliate of ARMA International, the Association of Records Managers and Administrators, www.ARMA.org).

Contact:

SteveGilheany@ArchiveBuilders.com
Tel: +1 310-937-7000 Fax: +1 310-937-7001

For more information, courses, and papers:

<http://www.ArchiveBuilders.com>.