

White Paper

Your Sarech Enigne Can't Raed This

By: Randy Van Ittersum and Erin Spalding CDIA+
www.disusa.com

OCR: A Complex Problem

The limitations of relying on text searching become clear when you use a search engine on OCR'd documents. OCR software has gotten much better, but you can count on 20+ errors on most non-laser-printed pages.



There is a little silliness out there on the web illustrating how you can still make sense of words if the first and last letters are intact, even if all the others are scrambled. That only works for HUMANS. Computers can only search the actual strings of letters.

To OCR A Document or To Label A Document

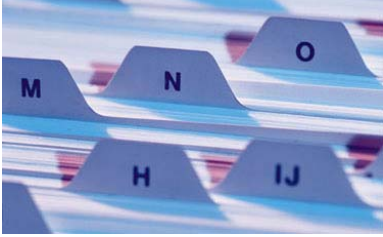
There are two significant issues that need to be addressed when making the decision about whether to OCR a document for search purposes:

1. Indexing the documents, and
2. OCR verses labeling the documents.

Both of these issues will have significant impact on which document imaging system to use.

Indexing the Documents

Before you can search for documents, you must first create an index from which to conduct your search. The time it takes to build an index is in direct proportion to the number of key search



words that it must include in the index. For example: a customer using our labeling system typically includes three key search words such as the customer name, number and date. A one-page OCR document may contain 650-700 words or over 200 times the amount of indexed words of a labeled document. If it was a 10-page document, it could contain 2000 times the number of indexed words of a labeled document. The result is that it could quickly explode the index so that it could take hours, if not days to build just one index.

To overcome the inherent limitations contained in the indexing process, a customer will have to seek out a more robust indexing system such as that offered by Google. Although Google is a solution to the problem, it's price starts at \$25,000. For a customer scanning 500-5,000 documents a day, this is not an effective solution because it costs \$25,000 compared to the \$0 to \$1,000 for other indexing systems.

OCR Verses Labeling the Documents

The decision to use OCR should revolve around the issue of data mining versus document retrieval. For documents and/or information contained within those documents to be searched, they must be indexed. The imaging industry is doing a disservice to its customers when it promotes only to OCR a document for search purposes, simply because it automates the process.



I believe the decision to use OCR needs to be made on a document-by-document basis. To understand my position, you must first understand the drawbacks associated with each OCR.

Search Results



While search engines (indices) are easy to use, the results they return are imprecise, all or nothing affairs. One can enter a word that describes what one is looking for, but one is then bombarded with endless lists of results that may or may not be relevant. For example, searching for “java” would return documents that describe java the programming language, java the coffee, and Java the Indonesian island. The greater the number of documents on your server, the greater the number of irrelevant search results.

Labeling documents will significantly improve the accuracy of your search and return a higher number of relevant documents. This is especially true for industries that are retrieving documents like job files, invoices, and standardized documents.

For other parties, like attorneys, it may be more practical to OCR the document because they may be searching for a key word or phrase contained within the document. Even then, an attorney should not just OCR every page contained within a document, but should select only those pages to OCR that would be relevant to a later search. For example, he may want to OCR only the page that contains the legal description on a real estate contract.

OCR is Not 100% Accurate:

Optical Character Recognition (OCR) is a process of converting text on a scanned image into text that can be searched. The idea then is to perform a “full-text search” on the OCR document with key words and phrases that are known to be included in a document. The OCR process is sensitive to the quality of the image as well as the differences in the fonts used within the

document. As a result, the output from an OCR process is seldom totally correct. If the OCR process claims to be 95% accurate, that means that one character in 20 is not recognized. Errors are introduced when characters “bleed” and touch each other or when the scanner picks up “ghost” images from the reverse side of a document. OCR software invariably substitutes “l” for “I” and “e” for “c”. Because of the inaccuracy inherent in the OCR process, it requires an operator to manually correct all the suspect characters.



A study was conducted by Imaging Magazine, which analyzed the cost, speed and accuracy of seven of the top scanners, quickly identified the accuracy problem of OCR.

Brand	Cost	Speed	OCR Accuracy
Ricoh	5,000	53 ppm	84%
Xerox	12,000	26 ppm	94%
Ricoh/Improvision	15,000	46 ppm	74%
Bell & Howell	20,000	46 ppm	84%
Fujitsu	26,000	89 ppm	79%
Banc Tec	65,000	144 ppm	84%
Scan-Optics	149,000	202 ppm	87%

If you are using OCR in place of labeling a document because the process is automated, the need to correct suspect characters negates any time savings gained by the automated process. And of course, you still have the problem of capturing an irrelevant OCR document in a search because the keyword is included within the document.

Fussy Search

I have encountered some companies that propose that it is not necessary to correct the suspect characters and you will usually find the document you are looking for. They suggest that you use “fussy search” technology. This technology expands queries to include terms that sound like or are typographically similar to the term requested. The problem with fussy searches is that they produce an even larger number of documents that are irrelevant.

When We Would Recommend to OCR a Document

I would recommend to OCR a document if you wish to conduct an intradocument search. This is especially useful if one is data mining or looking for information contained within a document. An example would be an attorney looking for a statement contained in a deposition.

One will navigate from occurrence to occurrence of the word, starting with the first “hit”. Each “hit” is highlighted within the document, thus making the search process much more efficient, and enabling users to retrieve and use information faster, in fewer steps.

Conclusion

Which process you use will depend on your own particular situation. We would suggest that you normally label a document, and when needed, only OCR those pages within the document that would be relevant to a future search. Because each word is entered into your index database, this method will keep your index database to a manageable size and return the most accurate search results.

Many people have found that it is faster and more accurate to manually enter predetermined keywords, phrases and numbers to a document than to OCR the document and correct the suspect characters. The flexibility of the labeling method is valued by people with a large number of documents that want accurate searches. The freedom to use characters and numbers when labeling documents allows operators to use existing terminology with which employees are already familiar. This facilitates a quick and easy transition into using digital documents in place of paper documents throughout their organizations.

Article by: Randy Van Ittersum, President of Document Imaging Solutions Inc. and Erin Spalding CDIA+, Instructor of CompTIA CDIA+ (Certified Document Imaging Architech), www.disusa.com , 616-847-5055