PRIMER

OCR, TEXT SEARCH and INDEX FIELD SEARCH

OCR Simplified:

Many years have passed since OCR was nothing more than a demonstration technology. It is being used successfully to automate data entry for many systems throughout the world. OCR creates ASCII text from document images by recognizing sets of pixels that form a single character. The number or density of the pixel groups on *an image is* controlled by the resolution of the *scanner.* The first OCR technologies relied on matching the exact pixels of each character to a memorized character set. In order to achieve greater accuracy, today's OCR technology deals with size independent and font insensitive features. The number of loops in the characters, the curvature, and other features are used to discriminate between difficult characters such as "t" and ''f," and "S" and '5."

In order for OCR to be effective at automating data entry, it needs to be accurate. Unfortunately, OCR accuracy is elusive. Changes that help in one case do not in another. A given change fixes one problem ***but creates*** another. Maximizing OCR accuracy requires persistence. You must be prepared to try several things in various combinations and must diligently test these combinations and track the results,

The most critical detail in maximizing OCR accuracy is also the most obvious. Make sure the original documents and the subsequent images are of the highest quality possible. Avoid faxes, dot-matrix print, and duplicate copies of originals. If you have control over the source document, try different fonts, font sizes, and line spacing. Make sure the text to be OCRed is well separated from other text. Try changing scanner settings to achieve the best possible scanned image and use a resolution of at least 240 dpi, preferably 300 dpi. Poor original documents or images will handicap your ability to maximize accuracy before you even start optimizing the OCR module settings.

Optimizing the OCR module settings involves changing the various settings to achieve the best accuracy. The details of these settings are beyond the scope of this paper. Establishing optimal OCR module settings can have a dramatic impact on the accuracy of the OCR.

Even after diligent testing you will find that OCR still makes mistakes. Since OCR is a repetitive computerized task, it will make the same mistakes over and over again.

Text Search:

After a document has been OCRed, a document index is created. The index is formed by parsing the document text into words. A sorted list of unique words is created for the document collection, and the location of each occurrence of a unique word is recorded.

When searching for a word, the search command is parsed into search words and operators. The occurrence list of each search word is combined with other occurrence lists using the operators, and the resulting occurrence list determines the qualifying folder/subfolder/document/pages/line/words, and the data segments are presented.

If word highlighting is implemented into the system, the location of pixels is also stored for each word.

Accuracy

A document text index is of little use if it's not accurate. Confidence in the system would be low because you could not be certain that all the documents matching a criterion were found. The best way to ensure an accurate index when using OCR is to validate the result. The best way to validate OCR is by comparing the results to a known good original or image. Persons looking at the index data and comparing the value to the image may do this manually. Even if manual data verification is performed only on low confidence documents, that can still be 5 percent, 10 percent, even 15 percent of the total. To make matters worse, just because OCR has a high confidence in its result does not mean it is correct.

Index Field Search

What is an index field?

When you go to a public library to look for a book, the first thing you do, is to go to the card catalog to search for the book and find out where it is located on the library shelves. You would have to decide whether you want to search using the book's title, the ISBN number, or the Author's name. These three search items are called index fields, and each card is called an index card.

Document Management Software will allow you to create as many index fields, as you want, on one electronic index card. You use this index card to find your documents, the same way you do at the library to find a book.

So, take some time, and think through how you want to index your document once it is scanned; and remember that index fields are data fields used to retrieve the documents, e.g. customer name, SS#, account #, etc. The rules for indexing are simple:

a. Use as minimum as possible of the index fields. One unique index is all you need to find the document (e.g. SS#, ISBN#, Parcel #, ten digit telephone #, etc)); if not possible, then the combination of two or more should be unique, but is not required.
b. If you can't form a unique combination, the system will pull all documents that meet the search criteria, and the user has to decide which one he is interested in.

The index field data is stored into a relational database engine, such as Microsoft Access, MS SQL server, etc. This feature facilitates the process of integrating document management systems with other enterprise information resources.

Accuracy

Using Index Fields is the industry-preferred method to search for documents. Data entry is accomplished using zone OCR to automate the process, and manual double or triple data entry to validate all suspected records; lookup databases are also used to validate standard fields such as cities, and states. The combination of these techniques makes this search method more reliable than the Text search described above.