

November 2008

Creating Effective Search Queries for Foreign- Language Documents

by Gary Wiener, Esq.

Executive Summary

Use of effective search queries can save a significant amount of time and money in electronically-stored information (ESI) document processing and review, by culling huge collections down to a more manageable size. Crafting effective word-search queries, though, can be troublesome, especially when the documents to be searched are in a foreign and unfamiliar language. The techniques used for creating effective foreign-language queries, for the most part, apply to creating English-language queries as well. Above all, the creation of queries must be treated as a fluid and evolving process, not just a one-time exercise. This process must be documented thoroughly, if one is to insulate themselves from a potential motion for sanctions.

Unlimited Storage, Unlimited Expense

At the local Peripherals-R-Us, I recently saw a one-terabyte hard drive on sale for \$139 – less than 14 cents a gigabyte. Indeed, storage media is now so cheap that people rarely delete anything from their computers. When the drive fills up, they buy another one.

This is not good news for those who have to comb through that terabyte of ESI in preparation for litigation discovery. According to the Sedona Conference, with billable rates for junior associates at many law firms now starting at over \$200 per hour, that single 14-cent gigabyte of storage can easily exceed \$30,000 to review.¹ Multiply that out to a terabyte – or larger – and the numbers can make even the wealthiest clients' heads spin. Mind you, that doesn't address the cost of processing that ESI into a reviewable, producible form. Can those costs be kept to a reasonable level?

The preferred method of keeping costs manageable is to run keyword search queries across the documents, in order to cull the mountain of data down to a much smaller size. Much has been written about concept searching as a preferred alternative to keyword Boolean and fuzzy searching. It's true that keyword searching is prone to return quite a few false "noise" hits. However, a closer look at the pro-concept search articles reveals that they tend to be white papers written by – yep – vendors of concept searching solutions.²

The fact remains that "the available evidence suggests that keyword and Boolean searches remain the state of the art and the most appropriate search technology for most

¹ Jason R. Baron, Ed., Search and Information Retrieval Methods, *The Sedona Conference Journal* 191, 192 (2007).

² An apparent exception would be Judge Facciola's opinion in *Equity Analytics LLC v. Lundin*, 248 F.R.D. 331 (D.D.C. 2008), in which he wrote that "determining whether a particular search methodology, such as keywords, will or will not be effective certainly requires knowledge beyond the ken of a lay person (and a lay lawyer) and requires expert testimony that meets the requirements of Rule 702 of the Federal Rules of Evidence." 248 F.R.D. at 333. However, the matter at issue in *Equity Analytics* centered on illegal access to electronic documents by a fired employee, and therefore the context of Judge Facciola's observation was most likely in the context of forensic examination of the questionable data, which requires specific expertise. See Debra R. Bernard and Mary Rose Hughes, "Much Ado About Text Searching", *National Law Journal* (Aug. 28, 2008).

cases.”³ Concept searching has simply not evolved to the point that it is a reliable replacement for keyword searching.

But let's say that the data is not in your native language, but in a foreign language that doesn't even use an alphabet that you're familiar with. What do you do now? How are we supposed to build effective search queries in an unfamiliar language when we face such challenges building them in our own?

Although this article will discuss effective search query strategy in the context of foreign-language documents, most of the techniques that we would use to build search queries in other languages are also effective to cull down English-language document collections. Word-search queries are rarely a perfect solution, but with careful planning, we can make them as effective as possible. Fortunately, the FRCP only requires a “reasonable inquiry”⁴ and “good faith” effort⁵ to find all responsive data in an ESI collection. Nowhere is perfection required – which is a good thing, because when it comes to ESI, there is no perfect search strategy. With careful planning and revision, though, it is possible to come reasonably close.

The Search Management Process: An Overview

With all of the various factors that must be taken into consideration (which we'll touch on shortly), there is really only one way to create effective search queries, in any language: Think of their creation as a process, not as a one-time exercise. Processes are fluid and adaptable to changing circumstances, whereas a single attempt at culling, a single time, is not likely to provide anything but frustration.

Under FRCP 26(g)(1), an attorney must certify that to the best of her “knowledge, information and belief, formed after a reasonable inquiry,” disclosures are “complete and correct” at the time they were made. For the most part, judges have not issued sanctions because disclosures were imperfect; they have sanctioned because the attorneys could not demonstrate good faith in producing those disclosures. Having a thorough, documented and (above all) good-faith process for culling responsive documents is always going to be the best defense to a motion for sanctions.

Consider the initial search and retrieval process a “pilot process”, which can be evaluated and modified as the review team learns more about the corpus of information from which to cull. One useful approach to test the process is to focus on a small subset of custodians at the center of the facts at issue in the case. These custodians, who are more likely to possess relevant documents, will provide a manageable sampling of data; you can then thoroughly review the results for responsiveness. This data will help you get a handle on the issues and terminology used by all custodians in the matter, and keep you on track.

³ See H. Christopher Boehning and Daniel J. Toal, “Assessing Alternative Search Methodologies,” *New York Law Journal* (April 22, 2008).

⁴ See, e.g., FED. R. CIV. P. 26(g)(1).

⁵ See, e.g., FED. R. CIV. P. 26(f), FED. R. CIV. P. 37(f). Keep in mind, though, that the court is not likely to take a claim of “good faith” strictly at face value. See, e.g., *Qualcomm Inc. v. Broadcom Corp.* (S.D. Cal. 05-CV-1968-B, Aug. 6, 2007) at 38, 51 (although Defendant claimed good faith and a reasonable search for responsive documents, they failed to produce over 200,000 pages of responsive ESI found with what the Court called “such an obvious search” query. The Court later sanctioned Defendant over \$8.5 million in legal expenses, and referred six attorneys to the California State Bar for disciplinary action).

Once you have your sampling, test different search queries across them, and rank their relative value based on how well they return relevant information with a minimum of “noise”. This may require the creation and testing of dozens of different queries, but have patience – extra time invested on this key sample data up front will save hours later on.

After the most effective queries have been identified through testing and ranking, they can be applied to the entire corpus of documents. You can do so with confidence that what you’ve done is consistent with the due diligence and good faith requirements of the discovery process and the FRCP.

We’ll discuss the management of the search process in greater detail near the end of this paper. For now, though, some basic familiarity with the way a search engine works, and factors to consider in the creation of foreign-language keyword search queries, will help you understand the process of effective keyword search query creation much more thoroughly.

Tokenization

When a computer indexes an English-language document, it “sees” everything between spaces or punctuation as a “token”. These tokens are catalogued in an index file, which contains a list of every indexed document in which that token appears. When you run a search for a particular token, the computer searches the index file rather than the entire document universe.

The same thing happens when foreign-language documents are tokenized, except that in many Far Eastern languages such as Chinese, Japanese and Korean (the so-called “CJK” languages), there are no spaces between words as there are in Western languages. Tokenization software customized to each language relies on pattern matching, so that tokens can be extracted to build the index.

When you want to search a foreign-language index, therefore, you must be careful that the elements of your search queries reflect the tokens as they are contained within the index. This can be daunting when you don’t know the language, and aren’t entirely sure where each word ends and the next one begins. Hiring a professional translator to build a foreign-language query is one solution; but once you know how the search function works, even someone who doesn’t know the language can piece together an effective query.

The Challenges of Chinese

As our example, we’ll use Chinese – certainly one of the most challenging of all languages for which to write search queries. There are actually two distinct flavors of the written language: “traditional Chinese” (Cantonese), which is used in Taiwan, Hong Kong and Macau; and “simplified Chinese” (Mandarin), which is used in Mainland China and Singapore. In simplified Chinese, about 30 percent of the commonly-used written characters from traditional Chinese have been changed. Since search engines are literal (remember, they rely on pattern matching), you must account for both possible versions in building your queries, unless you are absolutely certain of the origin of the documents to be searched.

Even so, prior to 1956, simplified Chinese didn't exist. Words that have been invented since then (such as "Internet") are therefore likely to be written completely differently between traditional and simplified Chinese. This is likely to be the case with numerous technical terms and concepts, particularly in patent documents (most of which are likely to come from the mainland). Therefore, be prepared to search for both versions of the key words.

Another factor to consider is that China has a number of different regional dialects. Foreign proper nouns are usually transliterated into Chinese by using written characters for their phonetic value, rather than for what they mean. Each Chinese-speaking region may (depending upon where the drafting attorney grew up) transliterate the same name very differently.

Compounding this is the Chinese use of acronyms, which we in the West build by taking the first Roman letter from each word in the phrase or name. In China, acronyms can be built by taking a character from each part of the word or phrase, but there are no clear rules as to whether these characters should be the first or subsequent characters in each word. This must be considered when you build your search query.

One Concept, Many Characters

In the CJK languages, the exact same word can be written a number of ways. In Japanese, for example, six different characters can represent the concept "sword". This is not so different from what we do with English (after all, "sword" means substantially the same thing as "épée", "saber", "foil", "rapier", or "scimitar"); but as with English, the synonyms have to be factored into your search.

The obvious problem is that more synonyms mean more false search hits. "Foil", for example, would not only return sword-specific uses of the word, but would probably return references to aluminum foil, comic foils, etc. "Rapier" might return references to "rapier wit"; "saber" might return hits for extinct tigers; and so on. Using Boolean "proximity" search techniques (example: "foil w/5 fencing" to find "foil" within five words of "fencing") would be an obvious way to limit these false hits, and searching in foreign languages is no different.

However, if a document has been scanned, the computer must "read" the document to make it searchable, by using optical character recognition (OCR) software. OCR software is far from perfect. It compares each individual character to its database of look-alike characters, and picks the one that appears to be the closest match. Even the best OCR software estimates only a 99 percent accuracy rate, which can mean that one double-spaced typewritten page – 60 characters wide by 25 lines deep – may have 15 misrecognized characters within the 1,500 on the page. If the document has a colored background, or has been faxed, or was scanned at too low of a resolution (a particular problem for Chinese characters with their tiny calligraphic strokes), the error rate will be even higher.

In English, we learned to read by recognizing words, and can therefore judge whether a misspelled or illegible word is close enough to the intended word to be understood. Computerized search engines have no such logic. If a character is wrong, the search engine will search for the wrong character. In Western languages (and even in Japanese), we can account for this by using "fuzzy" searching, at which we tell the search engine to look for tokens that have a certain percentage of matching characters. If we have a search that is 20 percent

“fuzzy”, the search engine will look for an 80-percent-accurate (say, four out of five characters) match.

More matches, though, means more false hits. “Sword” might return “swore”, “sworn”, “toward”, and so forth. The ratio of false hits increases tremendously as the search fuzziness increases. Even worse, fuzzy searching doesn’t really work for pictographic alphabets, such as Chinese. Because each character represents a concept, not necessarily a phonetic representation of a sound (as in English), one misinterpreted character can completely change the meaning of a sentence. Fuzzy searching in Chinese would return all sorts of false hits – and let’s not forget, the whole point of culling a set of documents with search terms is to cut down on the number of documents we must review.

Slang

Refer to any English-language thesaurus, and you’ll find very few words that don’t have a synonym. Nearly every word that conveys a significant concept has picked up not only dictionary words that mean the same thing, but also slang words and industry jargon that you may not find in the dictionary.

In the context of foreign-language legal documents, you must be especially careful to brainstorm as many possible iterations of industry-specific terminology as possible, and include those in your search queries. If you’re searching through e-mail (which tends to be written and sent without much thought given to careful grammar), you should also be aware whether your proper-language search queries have slang equivalents, and factor those words in as well.

For foreign-language synonyms of this sort, there is really no substitute for having a native speaker of the language (preferably somebody familiar with the specific industry at the core of the litigation), with whom you can consult for a list of slang terms and jargon that might help you find those elusive responsive documents.

Entering Search Terms

While a Chinese legal professional in Beijing might not have much difficulty typing Chinese characters into a search engine using a Chinese keyboard, a non-Chinese-speaking legal professional in San Francisco, using a QWERTY keyboard, has no such level of comfort. Depending upon one’s creativity, however, there are three useful methods for a non-native typist to load at least basic foreign-language terms into a search query.

The first is simply to find a word or name in a document that you are sure is written correctly, and copy and paste that word into the search engine. Nearly all search engines work by matching a sequence of characters to the search term, regardless of the language in which the characters are written. Bear in mind, though, that the search engine only does what it is told. A misspelled search term, or one on which OCR was performed (and characters misidentified), will return false “noise” hits. Also, entering search terms in this manner won’t account for synonyms, slang, or industry terms of art.

For traditional and simplified Chinese, an alternate method of entry is to use Pinyin, a method of “romanizing” phonetic Chinese pronunciation. This allows a typist to write Chinese characters using a Western-language keyboard, by typing in the Roman phonetic character,

then choosing from the appropriate Chinese character that appears on the computer screen. For a human translator, Pinyin is the preferred method of data entry. The obvious problem, though, is that Pinyin is useless for someone who doesn't speak Chinese.

A third option, while far from perfect, is Wikipedia (<http://en.wikipedia.org>). If an entry exists for the desired search term, the Wikipedia standard is to begin the entry with the subject appearing in boldface font. In the left-hand column of a Wikipedia page, you'll usually find a list of foreign languages in which an article on the same subject has been written. The foreign-language articles are almost never direct translations of the English-language Wikipedia page. However, the foreign-language page should also have the subject at the top of its entry, in bold-face type. While there has been much skepticism about the accuracy of Wikipedia as a research tool, the bold-faced word is almost certainly a direct translation of the one you're looking for. Simply copy and paste it into your search engine, and you have a correctly-spelled search term in the foreign language.

One tempting source that you should **not** use is any machine-based language translation software, whether as a stand-alone software package or over the Internet. Machine translation might be useful for giving a reviewer the gist of a document's contents, but it is all but useless for seeking an accurate translation of a single word. Unless that precise word is in the translation software's database, and unless there are no context-sensitive variations on that word, the software will return the "best guess" – which may not necessarily be the word you're looking for.

Search Query Techniques

Most legal professionals who have used an online case-searching service are familiar with Boolean searching, which applies the principles of algebra to the contents of an index. (A Venn diagram showing intersections and unions of the elements, which most of us learned in grade school, is an example of Boolean logic.) Boolean searching uses the connectors AND, OR and NOT, as well as proximity connectors (such as "w/5" meaning "within five words of").

Boolean searching is the most commonly-used technique for building word searches. Boolean logic lends itself equally well to foreign-language searches as to English-language searches, because the Boolean connectors operate on the tokens contained within the index, regardless of the characters contained within those tokens. So, if we wanted to search on six different ideographs for the word "sword", for example, we could enter the following Boolean query:

劍 OR 劍 OR 劍 OR 劍 OR 劍 OR 劍

Any document containing any of these six ideographs would then return a search hit. Don't overlook the NOT connector and algebraic use of parentheses to screen out likely false hits. An English-language example might be:

sword OR (saber NOT "saber-toothed") OR foil OR épée OR (rapier NOT "rapier wit")

We have already discussed the concept of fuzzy searching. In most Western languages (and, for that matter, in Japanese), fuzzy searching can be quite useful to broaden the range of possible search hits to be returned (at the expense of picking up a lot of irrelevant "noise" hits as

well). In Chinese, however, fuzzy searching is much less useful. Each character in Chinese represents a separate concept, which means there are few of them (if indeed, more than one) per word. Even a 20-percent fuzzy search, which would return a search hit for any token containing four out of five matching characters, would likely return a huge number of unresponsive hits.

Factor in, as well, the reality that so many Chinese characters vary by the minutest of calligraphic strokes, but those tiny variations have such wildly different meanings, that fuzzy searching is useless. Fuzzy searches work best where multiple characters usually combine to form a single token, so that a percentage of matching characters will return a reliable set of search hits with a minimum of “noise”.

Considerations for Other Languages: Japanese

The written Japanese language, while not as extensive as either flavor of Chinese, is nonetheless one of the most complicated languages. The principal reason is because there are four different alphabets (Japanese, Chinese, Roman, and Arabic numerals) that can appear in a Japanese sentence – indeed, within the *same* sentence.

The accuracy of a search using these characters is going to depend entirely upon the accuracy of the technique used to extract the text from these documents. If the processing software cannot recognize and change character sets on the fly, so that the Chinese as well as the Japanese and Roman characters are recognized, a significant chunk of potentially-relevant data is going to be missing from the search index.

Another factor to recognize is that Chinese characters used in Japanese writing rarely mean the same thing as the same characters mean in Chinese writing. Still another is that two words, written completely differently, can mean the same thing depending upon their context (for example, rice used for eating - 稲 - versus rice used for science - イネ).

As with Chinese, transliteration of personal names can be very tricky, as they may have a variety of pronunciations. Usually, personal names in Japanese writing are written using Chinese phonetic characters. If this name has been transliterated into English, the variations in pronunciation can cause a stick problem, of which the searcher must be aware.

Hebrew and Arabic

The two most widely-used languages in the Middle East, Hebrew and Arabic, are both written from right-to-left. This means very little in the context of searching, since most search engines don't care if the characters they're attempting to match are written from left-to-right or from right-to-left. The searcher does need to take care, however, to enter the characters in the same visual order in which they are expected to appear in the responsive search hits.

In contrast to the CJK languages, Hebrew has a blessedly-small number of characters in its alphabet (only 22). However, Hebrew words are usually written without vowels. This means that the number of homonyms – words that sound alike, but have different meanings – is much higher than you'd find in English. Many completely different words can be formed from a given root word. Prefixes are prevalent in most Hebrew writing, and in order to be effective, the search engine will need to be able to strip these prefixes.

Most of these considerations apply to Arabic as well, with one additional consideration. Arabic is a script language, and the characters tend to be joined. This is likely to have the greatest effect where the search is trying to cut and paste characters from a source document into the search engine. Careful character selection and entry, obviously, is critical.

Spanish, French and German

Even Western languages, which tokenize words much the same way as English does, require some forethought. Spanish, French and German use diacritics (marks placed above or below a letter or syllable to specify its distinctive sound value, such as accents, tildes and umlauts) that Americans may overlook. If the search index is accent-sensitive (meaning it treats a letter with a diacritic as a different character than its base letter), forgetting to include the diacritics in the keyword search may prevent a valid match. Fortunately, most search engines are accent-insensitive, meaning that a search term containing the letter “u” will also return search hits containing the letter “ü”.

Note, however, that diacritics may change the meaning of the word being searched, particularly if both the English and foreign-language version of the word are in the same index. For example, the English-language “mate” (which has numerous meanings having to do with association or associating) is very different from the Spanish-language word “maté” (meaning “I killed”). Similarly, the German brand name “Nestlé” has an entirely different meaning from the English word “nestle”. An accent-insensitive search index will pull up versions of the word with and without diacritics, meaning more noise to deal with.

The Importance of Testing the Keyword Search Process

In *Victor Stanley v. Creative Pipe*,⁶ Magistrate Judge Grimm threw out attorney-client privilege claims as to 165 inadvertently-produced documents. Even though counsel had applied 70 separate search queries to the document universe to filter out privileged documents, the court ruled that they had failed to conduct “quality assurance testing”. Clearly, making up queries off the top of one’s head – valid though they may be – is not an acceptable method of building a search methodology.

Begin by isolating the key custodians in the matter, and before running any search queries at all, isolate a statistically-significant sample (say, five or ten percent) of their documents.⁷ Review those documents for relevance and privilege, but note recurring keywords or themes in the responsive documents. Not only will this provide a useful benchmark against which to compare search results against the entire document corpus, but it will educate the legal team as to what to expect from the rest of the documents.

Once the sample documents have been isolated, testing of search queries can begin, based on the results of the manual review. How do the results of the search queries roughly

⁶ 250 F.R.D. 251 (D. Md. 2008).

⁷ See *Treppel v. Biovail Corp.*, 233 F.R.D. 363, 374 (S.D.N.Y.) (“[T]here is no obligation on the part of the responding party to examine every scrap of paper in its potentially voluminous files in order to comply with its discovery obligations. Rather, it must conduct a diligent search, which involves developing a reasonably comprehensive search strategy. Such a strategy might, for example, include identifying key employees and reviewing any of their files that are likely to be relevant to the claims in the litigation”).

compare to the number of times those results were found in the sample set? If there are too many “noise” hits, determine what might be an effective way to limit the noise. (Boolean searching allows for proximity searches, in which one word of the query must appear within a set number of words of a second word in the query in order to be considered a search hit.) Test a number of different keyword searches, and rank their efficiency (or deficiency).

Also, the resulting familiarity with different types of data sources will allow counsel to predict the costs of retrieval, backup and review of such data.

For example, the sampling process may reveal that certain data sources (such as local hard drives) or file types (such as Microsoft Access files) have such low yield that the collection and review effort is not worthwhile. While opposing counsel may not agree to such decision, the data provided by the sample review will provide the evidentiary support needed to defend the reasonableness of such steps to the court.⁸

After the most effective queries have been identified through testing and ranking, and you are confident that the search plan is defensible in court, you can apply the results to the entire corpus of documents. You can do so with confidence that what you’ve done is consistent with the due diligence and good faith requirements of the discovery process and the FRCP.

However, the single most important part of this process – regardless of the written languages involved – is the most frequently overlooked. Quite simply, you **must** document every single step of this process.⁹ By doing so, you can successfully defend your good-faith search process to the court. Without doing so, you may have no defense to sanctions at all.

Gary Wiener, Director of Litigation Services, is an attorney and electronic discovery consultant who focuses on developing effective strategies for managing electronic discovery, and helping clients use technology effectively in their law practices. Mr. Wiener is a member of the State Bar of Texas and a former Director of the Travis County Bar Association’s Young Lawyers Division. He is a frequent CLE author and presenter, and holds many electronic discovery industry certifications, including certification as an Oracle database administrator. He earned his degrees in Journalism and Law from the University of Texas at Austin.

⁸ Boehning and Toal, *supra* n. 3, at n. 8.

⁹ Even though a party’s search methodologies are arguably protected from disclosure by the attorney work-product privilege, in three recent cases, the court required attorneys to explain and defend to the court its methods. See *Victor Stanley*, *supra* n. 6, at 256; *U.S. v. O’Keefe*, 537 F. Supp. 2d 14, 23-24 (D.D.C. 2008); and *Equity Analytics, LLC v. Lundin*, 248 F.R.D. 331 (D.D.C. 2008). In *Victor Stanley*, the court found that by failing to disclose the keywords used to search for privileged documents, defense counsel had waived its privilege. *Victor Stanley*, *supra*.