H5

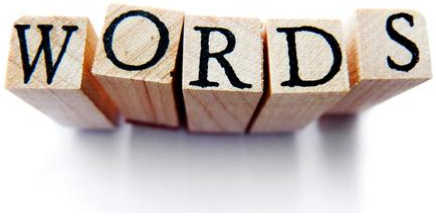
www.h5.com

Keywords:
Searching for “All and Only”

Improving the Art of Keyword Search

Keywords as Hypotheses

Keywords are used all the time as a means of trying to locate potentially relevant information in a document population. Creating a list shouldn’t be that difficult if you have some idea about the subject matter in question, right? But hazards exist in both directions: missing a keyword that may hold the key to what you need, and using keywords that bring in too much extraneous material. In litigation, the first is risky, the second, costly.



When you create a keyword list you’re essentially formulating an hypothesis—an “educated guess list” about how relevant subject matter will manifest in a particular document population. The more you know about the subject matter, the better the chance of finding what you need to find. Not surprisingly, whether or not all else is equal, your educated guess list may differ substantially from mine. Without question, however, both of our lists would certainly improve if we were to test them out and make modifications based on an analysis of the results. Sound scientific inquiry, after all, dictates that hypotheses be tested. Indeed, taking the time to iteratively test and analyze the results of a keyword list can yield great advantages: it can help reduce unnecessary volume in a review set (improve precision), and it can increase the likelihood that the resulting set actually contains the relevant documents required for production (improve recall).

Testing the Waters

Before you even create your keyword list, try to learn about any limitations in the indexing tool that has been used to process the population. Indexing tools have various parameter settings, and it pays to know if the one you’re using is going to cause a problem down the line. If numbers aren’t indexed, for example, the likelihood of finding that patent number you’re after will plunge to zero.

The first test you conduct will more than likely yield a few surprises. Keywords you thought sure to garner a large number of hits may not have, while other more questionable ones may have racked them up. Don’t accept these results at face value: investigate. You may find issues with the words you chose or the searches you constructed, or both. With proper analysis (and a bit of guidance), you’ll gain insight into how to tailor the list to achieve both better precision and

recall so that a re-run of a refined keyword list will return a more targeted result. Good recordkeeping is required during keyword validation: recording the results of individual keyword hits is a good way to memorialize the validation process as well as provide comparison points for future keyword tests over the same population.

Although it’s obviously best to review all of the hits, it is also acceptable to review a random sample instead in order to inform any changes to a particular keyword. While reviewing the hits, keep track of the approximate proportion of hits that seem to be on target. Precision and recall goals vary; if a keyword seems to be performing pretty well—say where 80% or more of the documents hit by the keyword were actually relevant—it is probably best left unchanged. However, should a keyword be performing at or below 50%—that is, it’s bringing in too much non-relevant material—it will likely benefit from revision.

Greatest Hits: Improving Precision

Although for litigation it may be safer to err on the side of over-tagging so that the odds of missing relevant documents are minimized, it is often quite easy to improve precision (i.e., eliminate false hits). From a cost perspective, it’s worth it. Most documents hit by a keyword search end up going through various levels of attorney review and those hourly fees can really add up. Any data reduction that can take place in advance of human review represents significant cost savings. There are several tactics you can use to improve precision.

Consider “hit volumes”

Start by reviewing the hit count of each individual keyword. Are there any keywords that have a much greater number of documents hit than you would have predicted? Keywords with a large number of hits are often the best places to look to improve precision as they are more likely to be over-inclusive, adding unnecessary and irrelevant document volume to the overall tagged total. An assessment of what got pulled in that doesn’t really belong there will inform how to refine the search to more effectively capture only what you need.

Add an anchor

Sometimes, a keyword needs an additional element in order to increase precision. For example, a keyword like “meeting minutes” might grab the targeted meeting minutes in addition to irrelevant meeting minutes. By adding a subject matter “anchor”—a keyword element that is thought to be highly correlated with the presence of relevant subject matter—precision can be

increased: “meeting minutes” could become “marketing board AND meeting minutes” or “marketing board w/10 meeting minutes.”

Use exclusion

If you can account for imprecision by identifying a particular responsible element, you can add exceptions to an existing keyword by using “NOT.” For example, “meeting minutes” could become “meeting minutes NOT executive committee meeting minutes.” This approach can be more time-consuming, however, as it often requires creating exceptions that are document-specific. Also, if a sample of a document population is being used to validate keywords, building document-specific exceptions is unlikely to address other irrelevant documents present in the document population but outside of the sample being used.

Revisit the use of wildcards

Wildcards, when used carefully with keywords, can safely increase recall and improve precision by covering variations of a concept. However, wildcards can also go haywire unexpectedly and the results need scrutiny to see if a revision makes sense. For example, if the original intent of the keyword “sting*” was to return discussions about stinging insects, you may not want those documents with the word “stingy.” Replacing the wildcard operator with a more limited set of keyword variants (“sting or stings or stinger or stinging”) or using an exception to exclude unwanted hits (“sting* NOT stingy”) can help to boost precision.

Use appropriate proximity operators

If a keyword includes a proximity operator, investigate whether reducing the operator size might result in increased precision. For example, the keyword “customer* and marketing” might be too broad, and could be replaced with the keyword “customer* w/25 marketing,” especially if you observe that the two keyword elements (“customer” and “marketing”) tend to be closer together in relevant documents than they are in non-relevant documents.

Scrutinize metadata

Sometimes, syntax allows the ability to draw on various metadata fields for use in keywords. When reviewing keyword hits, observe whether relevant documents tend to be within a certain date range or tend to be a certain kind of file extension. The related metadata fields can then be incorporated to refine keywords. For example, if a keyword seems to work only on documents with the .DOC file extension, add a metadata element to the keyword in order to limit the hits for that keyword to documents with this file extension. If a keyword seems to work in most documents, but also hits a number of

irrelevant .XLS spreadsheets, add an exception to the keyword to “NOT out” .XLS file extensions from the keyword hits.

After precision-focused refinements, it’s a good idea to run the newly refined keyword list and then look at the delta as compared to the original list in order to check whether any relevant documents are now being missed because of over-precision, and then make adjustments or additions as needed. There might be time and resources for only one round of precision refinements, but addressing potential sources of over-capture will end up saving time as well avoid any accusations of overproduction.

Missing Pieces: Improving Recall

Precision is only half of the picture, although it’s the easiest part to recognize. The results of high-precision keyword lists look really good because most of the hits are on-target. That said, this high-precision result is neither an accurate nor adequate representation of overall keyword performance, since looking at documents hit by keywords only supplies information about what kinds of documents are being hit, not missed. The part of the document population that has not been hit by keywords must be explored to make sure that the current keyword list is not missing relevant documents.

This is especially true for subject matter areas with a high degree of linguistic variability. The more ways a topic can be talked about, the more likely it is that you have not been able to anticipate all of the various ways to cover this topic with keywords. Perhaps even more irksome is that these missed relevant documents are not as intuitively on-topic as the ones being hit by the current keyword list: if these kinds of documents had been successfully anticipated, keywords would have been created to cover them. Untagged relevant documents signify that keywords have not been totally successful in tagging everything that should have been identified, which means production obligations have not been met.

Strategic search of untagged population

The first step in improving recall is to strategically search within the untagged population. Depending on the document volume and amount of time available, it might be possible to look through all of the remaining documents, but a random sample may make more sense. In order to increase the likelihood of finding remaining untagged relevant documents, broad search terms can be employed. Broad search terms target a more conceptual level of subject matter and will be less specific than individual keywords. For example, “marketing” is unlikely to be a

good keyword, but documents hit by this term that are not already being hit by more refined keywords would be a good place to start looking for missed relevant documents about more specific aspects of marketing. Other good broad search terms are similar to the kinds of keywords often intentionally avoided because of likely over-capture, such as abbreviations or numbers, as well as versions of existing keywords with larger operators.

Looking through the results of broad searches for missed relevant documents will clue you in to the types of keyword changes or additions needed. It might turn out that a proximity operator is too small, and that the operator size can be increased—or even changed to an “AND”—to improve recall. Review of untagged documents might also yield previously unanticipated synonyms and abbreviations that can be used to formulate new keywords. Often, documents with poor text quality are not hit by keywords and may require document-specific keywords in order to be tagged. As new untagged relevant documents are uncovered, looking for unique elements in these documents to use as keywords will help maintain precision while also improving recall.

Try, Try Again…

Whether because of time limitations or a lack of access to documents, it’s not always possible to test keyword hypotheses. Often, keyword lists must be formulated and finalized in a vacuum, without opportunity to validate and refine. In those situations, the hypothesis of the keyword list is never proven true or false—rather it remains a best guess. When iteration time and access are available, however, taking the time to properly validate keyword performance improves both the precision and recall of culling results. These improvements mean that legal teams are able to maximize their case preparation time by spending less time looking at irrelevant documents while making sure they have all the pieces of the puzzle to inform a legal narrative. It also makes keyword culling a more defensible process rather than mere guesswork. No keyword list is ever going to be perfect, but having a principled keyword validation process will help improve the quality of the output.

Tania Lihatsh, M.S.

Tania Lihatsh has worked in the legal services industry for more than 7 years. As a Senior Consultant at H5, Ms. Lihatsh supports H5 engagements by providing expertise in information retrieval and subject matter analysis, modeling and research, routinely advising clients in matters related to patent infringement, environmental remediation, and product liability.