

Document Review:

The Pros and Cons of Keyword Searching



Table of Contents

Keyword Searching	page 3
A Word about Precision and Recall	page 3
What's Wrong with Keyword Searches?	page 4
Not all Search Systems are Created Equally	page 5
What about the Leftovers?	page 5
What is to be Done?	page 6

The Pros and Cons of Keyword Searching

Document review is typically done by a staff of human reviewers reading and reviewing every document within the discoverable dataset to make determinations as to relevancy, privilege and substantive issues. The coders may be staff attorneys, temp attorneys or paralegals. As the volumes of discoverable data increases over time – mainly due to the proliferation of email - the practice of having humans put their eyes on every document is becoming time and cost prohibitive. In addition, there are many studies reporting the high-rate of inconsistencies found in human-coding trials.

One of the most commonly accepted methods for reducing the volume of the dataset to be reviewed is to develop a list of keywords and then run them as Boolean searches within the collection and review only the documents that were retrieved by those searches. Many of you are already familiar with the drawbacks of this method, and the findings of the well-known (and cited) Blair & Moron¹ research study which reported that these types of searches retrieved only 20% of the relevant documents (See discussion of recall and precision below) even when the attorneys involved *thought* they were retrieving 75% or more.

The bottom-line is that the document review process is ripe for an evolutionary change. By merging people, process and technology, document review can be completed faster, more accurate and cheaper than ever before. This evolution is already beginning to take place today and will continue into a mainstream methodology in the future.

KEYWORD SEARCHING

Many law firms follow the practice of culling the data collection through the use of keyword searches. In this scenario, the parties agree to a defined set of keywords that are then searched for, typically with a Boolean search engine, within the collection. The identification of and agreement on these keywords may be a long and arduous task between the parties involved. The new FRCP guidelines call for early 'meet & confer' sessions to discuss and agree on the handling of electronic documents and a considerable amount of time may be invested in this task.

Once the keywords are identified, typically the vendor doing the processing of the electronic data runs these searches in their system. These folks may or may not be skilled at forming constructive Boolean queries, or they may just feed in the keywords in a simple Boolean <OR> statement. Once the results are retrieved in this manner, they are 'tagged' or foldered in some way depending on the vendor and the review is set to begin. The next step and the most costly in terms of labor and time consumed is the task of having a human reviewer look at and analyze every single document returned in those results in order to make a determination as to the relevancy, privilege or issue.

The search and retrieval process narrows the overall volume of documents to be reviewed because it retrieves only documents that matched to the search or contained a 'hit'. The results of these searches typically reduce the overall volume of data anywhere from 10%-90% depending on the number and specificity of the keywords used. The comprehensiveness of these searches depends heavily upon the adequacy of the keyword terms and the structure of the search statements.

A WORD ABOUT PRECISION AND RECALL

There are two measures used in information theory that gauge the overall efficiency and effectiveness of search and retrieval systems. These measures are precision and recall. These benchmarks are in widespread use within the information retrieval industry and are standard measures for any type of search engine, categorization or classification system.

¹ Blair & Moron, "An Evaluation of Retrieval Effectiveness for a Fulltext Document Retrieval System", Communications of the ACM, Vol. 28 Number 3, March 1985

The Pros and Cons of Keyword Searching

Recall measures the completeness of the search – did your search retrieve ALL of the documents that are indeed relevant to your search statement. Precision measures the relevancy of the results – were the documents returned in your results indeed relevant to your query. An example may help us better understand the importance of these two measures.

Let us use the Enron dataset as an example. The dataset currently in the public domain has approximately 250,000 email documents in the collection. One of the issues relative to the Enron case concerns compensation received by certain Enron executives relative to the formation of the special entities they created to hide their losses.

We decide to use ‘compensation’ as our keyword term to identify the documents in the collection that are relevant to that issue. Let’s theorize that there are 3,500 documents in the collection that are relevant to that issue. When we run the search, we get 5,750 results and of those 5,750 results, 1,500 of them are actually relevant to the issue. Obviously, we have gotten documents that are not relevant but we have also missed some documents that are. This is how the recall and precision of that example would be measured:

Recall: $\frac{\# \text{ relevant documents retrieved}}{\# \text{ relevant documents in the collection}}$

Or, using our example: $\frac{1,500}{3,500} = 42\% \text{ Recall}$

Precision: $\frac{\# \text{ relevant documents retrieved}}{\text{Total \# of documents retrieved}}$

Or, using our example: $\frac{1,500}{5,750} = 26\% \text{ Precision}$

Those results are not very good and for any of you who have done keyword driven document reviews, you may know that it is not all that unusual. Our search term, compensation is overly broad and has multiple meanings. This volume of non-relevant documents (low precision) results in a tremendous

amount of wasted time by the review team – wasted time spent looking at documents from the HR director about a new sales compensation program, or a news story about Dallas Cowboy football players new pay structure rather than looking at documents relevant to the issue of the case.

In addition to the precision of keyword searches being low, the recall – or the ability to gather all of the relevant documents by our keyword search is also too low for comfort. We’ve missed over 2,000 documents that are relevant because they did not have the keyword “compensation” within the document.

WHAT’S WRONG WITH KEYWORD SEARCHES?

Nothing is wrong with keyword searches or Boolean search engines. It’s just that attorneys and even e-discovery vendors are not trained in how to use them or to understand the nuances of performing adequate searches. Often, we try to compensate (no pun intended) for these poor recall and precision results by expanding the search terms with synonyms to the keywords, so we might also search for pay, payment, fees, etc. – terms that are similar in meaning to compensation. This method will typically improve the overall recall – gathering up more of the relevant documents rather than missing 58% as in our example, but it may also lower the precision of the search because you have added in more terms expanding your results, but also capturing more non-relevant documents not focused on the specific issue.

Another technique used to improve the results of keyword searching is to use more complex Boolean search statements. For example, instead of just search for the word compensation, you might construct a search statement that combines other relevant concepts. For instance: compensation and (Fastow or Skilling or Lay), or possibly (compensation or pay or payment) w/10 (SPE). Both of these search statements are likely to improve the precision of the results because you have more than likely eliminated those articles discussing the Cowboys pay packages and HR discussions regarding compensation.

The Pros and Cons of Keyword Searching

While these methods are definitely an incremental improvement, there is still much room to improve the results on keyword searches and increase these two critical measures.

NOT ALL SEARCH SYSTEMS ARE CREATED EQUALLY

It is important to understand the nature of the search index created by the vendor performing the keyword searches. The search index is basically a catalog or table of the words contained in the documents in the dataset. Usually search systems use what is called an inverted index. An inverted index stores a cross-reference or mapping for each word found to all the documents where it is found and also tracks where that word appears in those documents. This type of index makes for faster retrieval of the documents when a search is entered into the system.

Not all indexes or Boolean search engines are created in the same manner and just slight differences can impact the results of your search. For instance, most search engines contain what are called stop words – these are common words that the search engine ignores. In some search systems, adverbs, adjectives and common terms such as computer are considered stop words. This may not pose an issue depending on your keywords, but could potentially result in missed documents. Usually, a system will allow you to force a stop word search through the use of quotation marks.

Another potential difference among various search systems is their stemming or truncation method. It is important to understand if the system does this automatically or whether you have to instruct it through your search query. For instance, if you search for the word compensate does the system automatically stem the word and retrieve all of its variations such as compensated, compensating and compensation? Or, do you have to ensure that you capture that through your search statement by searching for compensate*? It's also important to note what truncation mark is used by the system – is it an asterisk, a question mark or what? Different engines use different symbols.

Typos are a very common problem with keyword searching – just one letter difference and the word will not be identified. The very nature of email, its casualness and the almost rampant use of abbreviations and jargon make structure keyword searching even more inadequate. There are some systems out there now that allow for 'fuzzy' searching, basically allowing for one or two letters to be incorrect and this is an important advancement.

One final issue is that there is no standard within various search systems regarding search syntax or how proximity operators and/or Boolean connectors function. Most attorneys are familiar with search structure used in Lexis or Westlaw, but that is not necessarily how a document review platform will work. Some systems only allow basic Boolean connectors such as and, or, not while some allow for proximity such as mislead w/10 financials. Some systems allow you to search within the same sentence or the same paragraph – but many systems do not. It is essential to fully understand the capabilities of the system you are using in order to make the most of your search queries.

WHAT ABOUT THE LEFTOVERS?

One final concern with the existing practice is that during a document review rarely does anyone take the time to measure the recall or precision of the keyword searches. The time and effort involved in that type of effort is impractical as it would require a level of consistency and accuracy that is just not humanly possible. Usually, a high degree of faith is placed in the adequacy of the keyword searches and in the humans that are performing the review, but no measure of that accuracy is ever undertaken.

In addition, the documents that were excluded – the ones that did not match to any of the keyword searches are just ignored. No one really knows whether there are responsive documents in that set or not because they are not reviewed. Depending on the vendor of choice being used for the review, they may not even be loaded into the system. In addition to the possibility that they may be responsive documents, there may indeed be exculpatory documents in the set – ones that may help build your case. But these documents would have been discarded.

The Pros and Cons of Keyword Searching

WHAT IS TO BE DONE?

The intent is not to paint a picture of doom and gloom around the commonly accepted practice of Boolean keyword searching, but rather to highlight the necessity of using highly skilled and knowledgeable experts in forming the keyword lists, constructing the queries to be used and identifying the potential quirks and idiosyncrasies of the search engine being used. A better overall document review experience can be achieved if a methodology is used that combines the use of experts with the proper skill sets, a defined process taking into account recall and precision and the judicious execution of advanced search and classification technologies that are available today.

ABOUT RENEWDATA

RenewData is a trusted provider of services for the discovery, archiving, and governance of electronically stored information (ESI) to help organizations proactively manage the inherent risks associated with ESI.

RenewData's other offerings include e-discovery services that cover the five critical steps of the e-discovery process, including planning, preservation and collection, processing, review, and production. Our ESI risk management services provide corporations with a proactive means of managing the risks associated with ESI.

RenewData leverages unique and scalable technology, superior legal and technical expertise, and a high security facility to deliver defensible, accessible, and manageable data to our clients in a cost-effective and timely manner. RenewData has been ranked a top provider for four consecutive years in the Socha-Gelbmann Electronic Discovery Survey Report and included for three years in Inc. Magazine's list of fastest-growing privately held companies. For more information, visit www.renewdata.com or call 888.811.3789.



512.276.5500
888.811.3789

www.renewdata.com