THE SUFFOLK LAWYER
THE OFFICIAL PUBLICATION OF THE SUFFOLK COUNTY BAR ASSOCIATION
DEDICATED TO LEGAL EXCELLENCE SINCE 1908 www.scba.org Vol. 32, No. 6 – February 2018

INSIDE
FEBRUARY 2018
FOCUS ON
ADR

CYBER

Simple keyword searching is never enough

By Victor Yannacone

Many if not most e-discovery protocols are built around reaching agreement on keywords, but few of those protocols require testing to see whether the keywords are missing large numbers of relevant documents.

Precision vs. recall

Keyword advocates fervently hope that by achieving high *recall* — the percentage of relevant documents found; that *precision*— the number of relevant versus total documents — also stays high. It doesn't. There is a trade-off between recall and precision; the better the recall, the lower the precision. Unfortunately, the opposite is often true — better precision often means worse recall.

When the keywords seem to identify many relevant documents, the search seems precise, but with large datasets, there is almost certain failure to identify many other relevant documents.

Improving keyword searching

While keyword search can be effective in finding relevant documents, it can suffer from both low recall and poor precision.

Conducting broader searches to improve recall comes at the cost of lower precision and requires examining many more irrelevant documents. Total review costs go up accordingly.

Iterating keywords over a series of searches, sampling results and then refining the keyword searches can certainly improve results but only at greatly increased cost.

By relying on computer implemented algorithms, most of them proprietary trade secrets, to bring a desired level of recall while reviewing the fewest possible documents, commercial and academic predictive analytic systems often rely upon a process of weighing document features found through a continuous ranking process. Lawyers, paralegals, and other human beings are generally limited to identifying and locating the documents from which the searches are built.


Victor Yannacone

Basic preparation for keyword searching

If your client is required to produce documents, particularly ESI (electronically stored information), and you choose to use keywords and metadata features to create or limit a data set, at the very least you must statistically sample the data you are proposing to leave behind. Otherwise you have no viable defense against sanctions for overlooking what may be relevant and material documents.

In the search for appropriate keywords:

• Examine the complaint and every other document you have that is relevant to the subject matter of the litigation.
• It is best to start with broad terms. Recall is much more important than precision so err on the side of over-inclusion.
• Include word variants or "stems." Some review platforms have stem search capabilities, but it is best to think of all possible relevant variants and string them together with the Boolean OR operator. Beware of "wildcards," however, without some kind of preliminary testing for potential overinclusion.
• Include synonyms for any important terms. Check a good online thesaurus; but for industry-specific terms, check the trade publications, particularly those which have published "style manuals."
• Test, revise, and re-test your search terms. First run your searches individually or in small topically related groups. If the results demonstrate a need to revise, change only one element at each retest run. If necessary use well-established statistical sampling methods with robust randomization.
• Test *random* samples from both the document set created from keyword hits *and* the set of *all* the documents that have been discarded. Then compare the relevance rates of both sets before making permanent discards.

Remember, no matter how carefully you craft your search terms, keyword searching is imprecise. The only way to be sure your searches are sufficiently comprehensive is by testing your results.

Can a keyword search be as or more effective than with technology assisted review (TAR) and/or Predictive Analytics? That will be the subject of another, much longer column.

Note: Victor John Yannacone Jr. is an advocate, trial lawyer, and litigator practicing today in the manner of a British barrister by serving of counsel to attorneys and law firms locally and throughout the United States in complex matters. Mr. Yannacone has been continuously involved in computer science since the days of the first transistors in 1955 and actively involved in design, development, and management of relational databases. He pioneered in the development of environmental systems science and was a co-founder of the Environmental Defense Fund. He can be reached at (631) 475–0231, or vyannacone@yannalaw.com, and through his website https://yannalaw.com.