A Myth in the Making
Concept Searching vs. Artificial Intelligence vs. Human Review

by

Macyl A. Burke
ACT

I. Introduction

ACT is a litigation support company that works with large law firms and corporations in the discovery phase of litigation. Our services range from the collection of data to coding, culling, indexing, electronic discovery, including review for relevance and privilege to production.

This purpose of this paper is to offer a more complete explanation of filtering and reducing the size of large complex databases in the discovery phase of litigation. The paper also addresses the concept of “Artificial Intelligence”, or “AI” as some people refer to it and its impact on review for relevance and privilege.

For us the term AI is a bit like holding up a cross to a vampire. According to David Blair, a well respected information scientist at the University of Michigan, “Waiting for artificial intelligence is like waiting for Godot.” We agree with that opinion.

ACT has a great interest in reducing large document populations in a rational and useful way in the discovery phase of litigation and in speeding up the review process. Our learning in this complex area has been dynamic. Our primary educational resources have included a math professor from Caltech and other experts in the world of information science who represent organizations such as ASIS&T, AIIM, ARMA and The Sedona Conference among others. We participate with The Sedona Working Group Committee on Retrieval and Concept Search Applications, which is working to develop a set of Best Practices on search and retrieval.

Our raison d’etre is to reduce large document populations for review and then speed up that review and lower the cost. Concept searching may or may not be useful in that process. It is one of many tools that should be considered. Discovery is the largest cost in litigation (greater than the 80/20 rule) and review is the largest cost in discovery (greater than the 80/20 rule).

We can help our clients reduce large document populations. However, we do not believe Artificial Intelligence in the present stage of development is useful in this process. Bottom line, vendor claims have greatly exceeded their ability to deliver. Our advocacy is that process is more important than technology, and which technology is going to be most useful is situational.

The purpose of this paper is to open a wider dialogue on document culling and review between clients, potential clients, and interested parties. The immediate intent is to give a simple overview on some of the aspects entailed when employing concept search applications in trying to reduce large document populations prior to review.

There are two clear barriers to the use of concept search. We have an elephant in the living room regarding how competent people are at doing review. Missing from human review are metrics, measurements, and a set of best practices that been robustly tested. There is also a crazy grandmother lurking in the attic. The traditional economic model has been quite lucrative for law firms using partners. associates, paralegals, etc for review in large discovery cases. Associates doing large review projects have been big generators of revenue and some law firms are reluctant to give it up.

There is a conventional wisdom supported by some anecdotal information and the Blair Maron study that people may not be well suited to the review process and the congruence of agreement as to what is useful will vary greatly between individuals. At best the current information is inconclusive. No one I have spoken with who has made an attempt to more completely understand the dynamics of the review process, has been enthusiastic about what is being produced in the conventional manner which relies heavily on human activity.

The vast amount of the money of the legal spend is in review and that is where the struggle will take place. The first step has been the rise of contract attorneys who work at reduced hourly rates. Hourly rates are common in litigation but inadequate in and of themselves to produce useful metrics. A more complete set of useful metrics and best practices will have to be developed. The second step will be more sophisticated culling using concept search (rule based, statistical, and linguistic etc) coupled with statistical applications and sampling to provide boundaries on how well the populations have been reduced prior to review.

The name of the game is to reduce the population to a manageable size with metrics that verify the results and to then have humans review them for relevant, hot, and or privileged documents. The review needs to be measured with adequate quality applications to reduce variation in results.

Two overwhelming events have intervened into the discovery equation and are growing in influence. The first is the sheer size of electronic document populations that are increasing geometrically and are driving costs higher and higher. The second is the unwillingness of corporations to pay the rising cost of litigation without rethinking their approach. Many large corporations are following the path of DuPont and GE who are world class incubators for best practices and aggressively managing their legal costs like any other part of their business.

II. Searching & Culling

We are all familiar with keyword searching. If nowhere else, Google can provide the experience. There are problems using keyword searching for selecting useful documents in the discovery process in litigation. There are also places where keyword searching can be useful. However, when trying to select a useful set of documents from a larger population it can be very problematic. Keyword searching is over inclusive in what it selects and it leaves behind a high percentage of relevant documents. There is good empirical evidence to support this statement and the validity of the problem is widely accepted by experts in the field.

Concept searching is thought to be a better and more useful application. The evidence would support this contention. However, the discussion is badly blurred by the hype and shrill claims of a number of vendors hawking their wares.

To talk with intelligence about concept searching requires some balance, moderation, and a dose of skepticism. There is no Holy Grail. However, like Percival in his quest for the Grail, we will definitely be searching. It is a quest shrouded in a great deal of myth. And like Percival our learning will come from the journey and not the destination.

For ACT, concept search is not a technology or software application. It is a process, if done properly, with multiple steps. One size does not fit all. It is not a black and white approach. It can be simple or complex in application depending on the project. There is no magic bullet. Concept search applications can be useful or not useful depending on the need and requirements.

There is one overriding theme that we feel is key in regards to concept searching. It is the process that is important. It is not the tool or technology.

There are multiple players in the field of concept search. They include, but are not limited to, ACT, H5, Arête, Cataphora, Syngence, DolphinSearch, Engenium, Attenex, Autonomy and Meaning Master. Some sell direct, some sell through distributors, others sell engines, and some take work in-house and process it.

III. Basic Approaches to Searching

There are three basic theoretical approaches to concept searching:

1) Rule Based
2) Statistical
3) Linguistic

These might overlap in some applications, or be mixed in approach, but broadly they fall into the three basic categories listed above. There are advantages and disadvantages depending on the situation to which one should be deployed. One is not overall superior or inferior on the surface. Again it will be situational as to which would be the most useful.

The controlling element for selection of an approach is the nature of the population and the need and resources of the client. The approach is a bit like a doctor with a patient who walks in who has never been seen before. There has to be some diagnostic work done before recommending a solution. If the person is gushing blood from a severed artery, we would take a very different approach than we would towards a patient that complained of sharp stomach pains. The latter may be in much worse shape than the former. We will have to dig a bit to determine what is in front of us to determine the best course of action.

1) Rule Based Approach:

A rule based approach is sometimes referred to as a taxonomy or thesaurus solution. Taxonomy and thesaurus are sometimes used interchangeably, but they are different terms of art to professionals. The differences are subtle and for our purposes they can be assumed to be the same. A rule based application is an “if this, than that” approach.

With the word “pitch” you would retrieve general categories such as baseball, roofing, sales, etc. If you next selected “baseball” you would get fastball, curve, throw, etc. and on and on clustered around baseball. The same word could apply to horseshoes. The categorization is what is important. The sales category would bring up presentation, etc. It is hierarchal in nature. All of this indexing would have to be created around a set of rules that would then be used to tease out useful data.

It is labor intensive to construct.

2) Statistical Approach:

Statistical applications use mathematical applications to find data. The most common is Bayesian Statistics. The statistical term of art that Bayes’ theorem is working from is “conditional probability”. Very simply stated, “If I know that A has occurred, then, on those trials of the experiment where A has occurred, how often does B occur?” This statistical application lets you locate and rank related topics using probability. This model typically not only allows an initial relevance ranking to be more accurate but it also provides a mechanism for iterative searching based on relevance feedback. It is beyond our interest as to how this works in specific. Our interest is simply to differentiate the statistical approach as a method.

Statistical applications are mathematical and can be deployed quickly. The construction of the applications is part science and part art. Results can vary.

3) Linguistic Approach:

Linguistic applications, in addition to focusing on context and circumstances of use in language, look at the practices and activities in which the words and phrases are used. Linguistic applications rely on statistical applications, but require in addition deep analysis and inquiry of the culture and activities within the environment that the language is being used.

This can be time consuming in the set up or modification of any given project.

For our purposes, it is enough to know there are three basic theoretical approaches to concept searching based on different premises. They are rule based, statistical, and linguistic. They have different strengths and weaknesses. Circumstance would dictate which would be the most useful in any given situation.

IV. A Modified Statistical Approach

We use Autonomy, a statistically based Bayesian approach, which was judged to be pragmatic. It does not mean we would not recommend a rule based or linguistic application if that were in the best interest of the client. Autonomy was selected in the way a carpenter’s first tool purchase might be a good framing hammer. We may very well add other tools or applications to our in-house capabilities at a later date. There are three things we need to consider concerning any given population regarding concept searching:

1) Richness of the population in terms of relative or useful documents
2) Size of the population
3) Complexity and number of issues

The richness is the percentage of the population that is useful or relevant to the question or issue being searched for. This can be determined by taking a random sample of the population and analyzing it. This will give information on how many useful documents we can expect to find. As we know, a statistically valid random sample can be drawn from around a 1000 voters that will accurately predict the outcome of tens of millions of votes cast.

If we took a statistically valid random sample of 1,000,000 documents and the sample showed that 10% of the sample was useful or relevant, we could extrapolate that we would expect the population to yield around a 100,000 good, useful, or relevant documents plus or minus some margin of error. We would have some idea of the order of magnitude of what we are dealing with and the percentage of richness we could expect. A richness of less than one half of one percent could be problematic. The higher the richness the better results we can expect.

Once we have selected a sample and know what to expect as to the quantity in our target population we could use other statistical applications to give us further information. We could use sampling techniques to test the retrieved documents from a population for the confidence levels of good documents we got and the confidence levels of good documents we left behind in the original population to see how effective our efforts were in finding the useful material.

How the statistics are employed is not important for our purposes. What is important is that we can use metrics to quantify, qualify, and give objectivity to our results. All of the sampling results will be expressed in a range of plus or minus confidence limits. As an example, from a sampling result we might say, this population contains 46% percent relevant documents with a 3% plus or minus margin of error. Statistics and sampling can be powerful tools to understand the order of magnitude of a large and complex set of documents. With all measurement there is "variation". The margin of error is helping us understand that. We need to be thoughtful about tolerances and variation. The standard for discovery is not perfect discovery. It is reasonable discovery.

V. Use of Statistics

Some Federal Judges accept that a more reasonable way to review a document population of millions of pages is through the use of statistically valid applications rather than have humans do it. This is particularly true in construction defect cases. There is no settled precedent with bright lines, but the trends favor it. Statistics are being used and accepted now by courts, but it is not in universal use without question.

Concept searching is affected by the size of the population involved. The larger the population, the more problematic. There are ways around this (break it into smaller sections, etc), but size does matter.

The complexity and number of issues in a population matter. It would be impossible to look for 500 issues. Looking for a complex issue like round trip transactions is problematic. Round trip transactions are usually found by auditors after a company collapses. It is my view, round trip transactions would best be found by keyword searching and other types of filters targeted by the evidence uncovered in analyzing the collapse. The issue is too complex for concept searching, in my opinion.

There are two retrieval terms of art in information retrieval we should be familiar with:

1) Precision
2) Recall

Precision is a measure of how many documents in the retrieved set are relevant, with high precision indicating most of the documents returned are relevant. High precision would mean that most of the documents in a set are good.

Recall is a measure of how many documents that are relevant get found, with high recall indicating that most of the relevant or good documents got found. An example of the difference between the two measures follows:

Assume a data base of 1000 documents, of which 100 are truly responsive to a discovery request. The search retrieves 200 documents, but of the 200 “hits” only 50 turn out to be responsive.

Recall rate = 50/100 = 50%

Precision = 50/200 = 25%

We could then describe the retrieval with the following metrics. The retrieval population had a recall of 50% and precision of 25%. This would mean of the 100 good documents in the population we got 50 and left 50 behind. Of the 200 hits only 50, or 25% were useful and that 150 or 75% were not useful.

A perfect retrieval would be a result of 100% precision and 100% recall. That would mean every document we retrieved (100 in this example) was relevant and every relevant document was found (100 in this example). That would be the Holy Grail. We will not get there, but we will try to get as close as possible. Very high percentages of recall and precision should be viewed with suspicion.

Traditional search technologies usually offer high recall or high precession, but usually not both. Our goal is to use our process and application expertise to strike a good balance between recall and precision. Again it is the process that is key and understanding the nature of the problem and needs of the client and not the technology.

We are looking for a few needles in the haystack. The best way to do that is to remove as much of the haystack as possible before looking for the needles. There are a vast number of ways to do this (triage, de-duplication, file type filter, date filter, keyword, concept search, etc.). Once we determine that we are ready to employ concept search techniques based on the need, we would select a rule based, statistical, or linguistic approach as indicated by circumstance.

VI. Conclusion

We should be agnostics about tools and technology and ideologues about process.
The needs of the project dictate our solution to the problem, and not the restraints or limits imposed by a specific technology.

Currently there is not a single paper with empirical evidence published on concept searching. ACT is collecting data and will attempt to publish such a paper in the future. We are exploring this possibility with various clients and potential clients currently. Our intent is to do this with outside independent experts to increase validity. We welcome other interested parties to the fray.

The Grail in myth is less about an object than an idea or state of mind. The journey seeking the grail is the learning experience and the real value. The finding of it is a vanishing horizon. In terms of concept searching we will always be learning and never become learned. Honor the person who seeks the truth and beware the person who has found it.

May 16, 2007