AWAY WITH WORDS: The Myths and Misnomers of Conventional Search Strategies and the Search for Meaning in eDiscovery.

By Thomas I. Barnett, Special Counsel, eDiscovery and Data Science, Paul Hastings LLP

Away with words

You are in a highly contentious litigation, taking the deposition of someone you believe to have the most important information you need in order to win the case. It's the moment of truth, and, handing them a list several pages long, you ask, "Please tell me every time in the last 5 years you have used the words on the list I just handed you."

Does that scenario sound absurd? It should. But it is a fairly good approximation of the legal profession's overuse and over-reliance on key word searches to identify relevant information in large data sets. And, for better or worse, predictive coding, the latest, and some would argue greatest, approach to reviewing documents, is nothing more than key word searching on steroids—comparing, matching and ranking the entire set of words in a document with those of other documents in the set—based on the frequency and proximity of the words contained in the document. Fortunately, there are other ways to attack the problem thanks to some recent advances in technology and data science.

The approach of using key words, or any of their variants currently offered in the eDiscovery industry (e.g., "concept" search, clustering, and predictive coding) is based on what works well for computers—not on how people actually think, learn and communicate. Computers are far superior in speed and accuracy in identifying patterns (in 1's and 0's) and matching them and they get faster every year.

The foundational technology used in most currently available approaches to searching, identifying and classifying discovery data has been around and in use in the world of computer science for decades. The only novelty, to the extent any exists, is the relatively recent adoption of such common place technology in the legal field.

Specifically, most technology currently available in the legal industry to classify documents for responsiveness or for issues in a matter relies exclusively on the specific text of the documents—either individually or small groups of words as in key word searching, or grouping some subset or all of the words in an entire document as in predictive coding.

What these approaches fail to do is incorporate and correlate various types of data that can be easily extracted and brought into the process. Doing so brings us closer to the "*who*, *what*, *when*, *where*, *why* and *how"* questions that we actually use to understand events.

The fastest most powerful computers in the world can't even come close to our ability to consider the context of events or communications, determine and rank the importance of statements or actions, or interpret nuance in language. These are the very things that allow us to understand and make judgments every day and to arrive at what we believe to be the actual *meaning* of events, communications, and documents. There are clues to the context and meaning of documents that have yet to be fully exploited in the analysis of potential evidence in litigation fact investigation and discovery.

Cracking the semantic code

How might a skilled lawyer get the required information from the deposition witness in the above example? You might try to establish some background, foundation and context for the questions you want to ask. You might try to establish the timeframe of the events at issue, who said what, when, where, and, if possible, why certain things happened the way they did. You might seek to learn how people communicated about specific events—not by guessing what words they might have used, but by assessing the substance or meaning of the communications from many different perspectives. That way of thinking is basic to how we communicate, question and learn about the world as human beings. It so instinctive you probably don't even stop to think about it.

In fact, there is a tremendous amount of information that forms our perception of the world and how it works beyond simply the specific words we use. Consider this example from an old joke: you are a tourist in London and you stop and ask a policeman, "Do you know how to get to the London Bridge?" He replies, "Yes, of course," and promptly turns and walks away. We all know that you were asking for directions.

But at the basic level of most commonly used approaches to find relevant documents in litigation and regulatory matters, such a nuance would never be detected. Because when it comes to trying to find meaningful information about events from the text of documents, we as a profession have come to rely on what can be highly unreliable methods of gathering knowledge from potential evidence in our cases.

In recent years, however, unprecedented advances in a number of areas of computer and data science have opened the door to better ways of finding information and making sense of it. Two important such advances include (i) the advent of much cheaper and more flexible technology infrastructure leading to the refinement and wider use of so-called "parallel processing," using groups of relatively inexpensive computers together to achieve processing speeds and power exceeding anything we have seen before; and (ii) the advancements in machine learning and statistical engineering allowing us to tackle problems in a more intuitive, human-like, way. These advancements are opening new avenues in fact investigation and document discovery as well.

The predominant technology used in the legal industry to attempt to derive information from large sets of documents and communications relies on a very limited set of parameters. The methods are generally highly static and rigid.

A simple example is indexing the text of documents that need to be reviewed. In the eDiscovery industry text is typically indexed in a specific way by well-established tools either generally available or proprietary to particular providers. Almost invariably, certain types of characters are thrown out automatically because they are presumed to carry little or no semantic content ("stop" or "noise" words). Examples include punctuation marks, symbols, single letters and numeric characters.

But what if the substance of the case relates to the use of symbols, letters, or numbers such as a case involving mathematical formulas forming the basis of financial decisions or chemical formulas for pharmaceutical products? You will likely be out of luck in fine tuning based on the context of your case.

This is just a relatively minor tip of a much larger iceberg. There are subtle patterns and dynamic structures in communications that go far beyond the standard approaches to grouping documents whether by key words, clustering, proximity, co-occurrence, and word frequency. These approaches also go beyond applying the same document-level machine learning classification algorithm to each data set regardless of the unique attributes of the facts and issues in the case and the data set itself as in most "predictive coding" offerings.

A more substantive example than indexing is the influence on the meaning of communications based on semantic context and combinations of words. If your case involves allegations about accounting fraud you might be interested in improper reporting of financial transactions. Colloquially, some obvious phrases might include "cooking the books," "massaging the numbers," or "inflating results." Of course, none of these phrases may appear anywhere in the documents and communications. "Cooking," "massaging," or "inflating" can appear in all sorts of contexts that have nothing to do with the allegations in the case. You could try looking for "cooking" within some number of words with "books," but that will only work if you have somehow guessed exactly how the people in your case discuss this issue out of the virtually infinite possibilities of human language.

Word searches or their variants will likely be vastly over-inclusive and at the same time may miss the critical communications that were not divined in a key word creation exercise. All of this doesn't even consider that fact that, in the context of committing fraud, people often try to obscure their communications about activities they believe to be improper or illegal.

More things in heaven and earth

There are patterns and structures that exist in data that have subtleties and nuances well beyond the level of discernment of the highly standardized approaches to organizing data commonly used in the legal industry. The question shouldn't be how, generally, people may refer to improper accounting activity. The question should be how did these specific people communicate about the specific issues in this case even absent any idea about the precise words they might use.

The ability to do more subtle, flexible analysis exists—it is just not part of the mainstream of techniques used by most attorneys nor offered by most service providers. Such processes and technology has been developed by businesses and in academic circles and they have been applied successfully in industries that relay on high speed analyze of vast volumes of textual data.

In addition to identifying more subtle structures and patterns in text, data sets contain additional information that can be incorporated into the analysis. While these data components are familiar in the legal industry, they are not typically combined and analyzed together in attempting to derive knowledge and information beyond whether certain words can be found in a set of documents and communication:

> *Metadata* is information that accompanies our documents (e.g., emails, word processing documents) but is not part of what we think of as the content of the document. Such information includes the date and time the document was created, accessed, modified, or the time an email was sent, received, opened, and who sent it and received it. While such data is commonly used to search and sort data, it is not as typically combined with other types of data to determine meaning. Further, the

more complex approaches like predictive coding and concept searching generally rely on the text of the documents alone.

Entity data (also referred to as *extracted* metadata) is information contained in the text of documents that can be identified and classified into pre-defined categories such as names of persons, organizations, locations, quantities, monetary values, and so on. In other words, instead of looking at the text of a document solely as a string of characters that may or may not match another string of characters, this approach incorporates knowledge about elements of the document into the analysis. What do the words in the document actually refer to or mean?

Social network analysis refers to a way of analyzing what people are saying about a given topic, when they are saying it and who is saying it. As in the above examples, this type of data is often collected and can be analyzed using existing eDiscovery tools. But it is not typically analyzed in conjunction and aggregated with all of the other data attributes and textual analysis to arrive at a more case specific understanding of the communication patterns and structures that semantically define the data set in the case.

By way of example, suppose you have an email from one person to another with text that matches a set of search terms. That by itself may not provide enough useful information to arrive at a meaningful conclusion about events in the case. But what if you were able search for documents based on whether they discussed specific events in a variety of ways, using different expressions, at specific times, involving specific people? That is possible when using metadata and entity data along with text and social networking analysis. This is just one example of what is possible by incorporating and blending technology and processes used in areas outside of the legal profession.

We need to get beyond the confines of conventional approaches. We rely on these approaches either out of habit and familiarity, or because these are the only offerings being plied by providers which have invested in their development. In order to survive we need to take advantage of the best available techniques and technology to get at the meaning of the data not just simple pattern matching. Failing to do so will leave us drowning in the ever increasing ocean of data.