

Better Needle-in-the-Haystack Searches

A new generation of document-search technologies
will ease the task of discovery

KROLL ONTRACK®

TRIAL GRAPHIX.

When a company receives a document request requiring the collection and review of millions of pages of potentially relevant material, the shiver of fear that runs down the collective spine of the organization can be overwhelming. And when that material is due immediately, fear can turn into a feeling of chaos.

Chaos, in other words, is caused by that expedited request for a terabyte of data, involving millions of pages of documents, images, and various other media maintained, sometimes, by scores of employees in locations around the globe. But order can come from reducing that dataset – i.e., the compilation of all potentially discoverable material – to the 5 percent to 10 percent of information that is actually relevant.

The current evolution of that chaos into order is a direct result of the developments in “ranking engine” and “concept search” technologies that have swept the discovery landscape over the past few years. Ranking engines match keywords with material in a data set, while concept-search technology tries to locate items with similar meanings, rather than those with identical terms, so that reviewers can find out what the documents mean instead of just what they say.

This is a dramatic departure from traditional review methods and could mean millions of fewer pages to process, saving the client millions of dollars.

Traditional Review

Conventional document discovery required photocopying, Bates stamping, and review before anyone could even make an actual determination of responsiveness or privilege. Assuming that an associate could thoroughly review five boxes of discovery materials per day, it would take teams of lawyers years to get through a single review only to realize that a fraction of what they had seen is actually relevant. Furthermore, because of the volume and sheer confusion involved, the senior lawyers on the case sometimes never saw the most important documents or knew whether they had been missed.

Today, traditional review has been almost completely abandoned. The strategy for reducing this type of dataset was simply to collect as little data as possible in an organized and efficient manner while still complying with the requirements of the document demand. This shifted, rather than reduced, the chaos of discovery from the collection process to management of the material,

only making surprises at deposition or trial a greater fear.

Databases to the Rescue

With the advent of document-management databases, the chaos is further reduced. These systems house the complete metadata and files yielded from the electronic-discovery process, as well as images of hard-copy documents, which are collected, scanned, and coded to extract relevant keywords that can be easily searched and organized. These databases allow reviewers to analyze the material once, and make the generation of subsequent copies less costly because they can be automatically renumbered and reorganized. They also permit attorneys and paralegals to record their impressions, opinions, and subjective determinations about the documents; reduce the risk of losing track of key documents; and enable chronological ordering of material using rudimentary search functions that can identify Bates numbers or points of collection.

Despite these advantages, a database does not actually reduce the number of documents that must be reviewed, nor does it highlight those documents most likely to be relevant to the case. Lawyers are still reviewing millions of documents but are doing it in a more organized fashion, using a centralized repository that allows them to evaluate with an eye toward a more detailed secondary review by senior members of the team

The Next Generation

The goal, of course, is to eliminate the need for that secondary review altogether. By complementing human analysis with mechanical intelligence, document-discovery tams can achieve much higher levels of efficiency, which often translates into significant cost savings on legal fees.

Current technology that uses ranking engines and concept search terms can organize and prioritize data before an attorney handles a single piece of paper. It can effectively take millions of documents and divide them into subjects, time periods, and locations, and produce the few hundred thousand that actually need to be reviewed. It will also let the initial evaluation team decide which documents should be reviewed by whom, whether by contract lawyers, associates, or partners. Though these programs are most effective with electronic files, they can

also be used on hard-copy documents, which are then scanned and coded so that the key search terms will still be recognizable.

Ranking Engines

Ranking engines operate like Google but search only the dataset, rather than the entire collection of information available on the Internet. They then position results based on how closely they mirror designated search terms.

Though ranking has organizational advantages, users must understand that the algorithms work according to protocols implemented during the initial design by the developer and, therefore, many not be optimally successful. Also, the title of a document and word appearing at the beginning of a sentence, or those that are repeated frequently, tend to be given more weight, which could distort or confuse a search.

Some ranking engines even make distinctions between upper and lowercase words, or can mistake words that have similar spellings but completely different meanings, such as "hard driver" (which might be used in an e-mail describing a supervisor) and "hard drive." In a related issue, the software may not be adequately analyze stemmed words like "big" the "biggest." And, of course, ranking engines will generally not provide results using words with similar meanings, such as "contact" and "agreement."

Despite these flaws, ranking provides structure to a dataset and lets users prioritize the data so that senior attorneys can review it.

Since most collected documents have little or no true relevance to the case, concept search engines supplement ranking engines with linguistic logic to analyze a dataset. They look at the topics or concepts that recur in the text surrounding the original search request and make suggestions about other ideas that may be relevant to the reviewer. At the most basic level, concept search engines create relationships among the words found in large sets of documents, building a type of numeric characterization.

For instance, if a team reviewing antitrust documents enters the query "We are going to eat X company for lunch," the system would likely understand there is a greater meaning in this phrase and find other similar instances if business-strategy commentary, rather than search for issues related to meals and food.

This technology is useful for identifying the entire population of individuals in an organization who may be involved in the subject dispute, rather than those listed in a document demand. It could also reveal certain documents maintained by these individuals in a fashion that deviates from corporate protocol. Finally, analyzing concepts this way could confirm certain assumptions or even reveal critical answers a party to the litigation may be looking for.

While sophisticated ranking engines may allow advanced searches using "and," "or," and other connecting variables, rather than single terms, they generally do not allow natural language queries. Concept search tools permit free-form queries or requests for entire portions of a key document (like the requirements of a court-ordered discovery request), letting one know the difference between the keywords found and the concept-oriented results. Some packages feature "neural networks," which use a type of artificial intelligence to recognize user preferences based on a combination of prior use and examples built into the system. This is useful for managing larger productions being processed over time.

Unlike previously used techniques in traditional document review, concept searches can sample the work product of the lawyers involved and measure the accuracy of their review. It also enable the reviewer to sort results graphically by folders, search terms, and concepts. Complex mapping systems collect documents, e-mails, and other materials into groups and match with various items that have similar meanings. Not only does this feature make the review easier, but it also makes it more intuitive and, ultimately, more efficient.

Millions of documents cannot be reviewed in an expedited fashion without a little chaos, but eliminating almost all of it as non responsive before anyone lays eyes on a single page will never occur unless that chaos evolves into order. The ability to create order requires both an understanding of developing technology and when it can best be used.

©2006

ALM Properties Inc. All rights reserved. This article is reprinted with permission from LegalTimes(1-800-933-4317. LTsubscribe@alm.com. www.legaltimes.com)