MAYER·BROWN

Electronic Discovery & Records Management

TIP OF THE MONTH

May 2010

How to Manage the Risks and Costs Associated with Searching ESI

Scenario

A large corporation is served with a complaint accusing it of participating in a price-fixing conspiracy. Multiple discovery requests follow seeking electronically stored information (ESI). In-house counsel speaks with the company's IT department to estimate the scale of the review and is disturbed by the sheer number of files to be reviewed. How can a meaningful review be accomplished in a reasonable time frame in a cost-effective way?

The Risks Associated with Using Search Terms

The use of search terms has become the new panacea of many electronic discovery vendors that trumpet the use of technology to reduce the costs of identifying relevant documents and the number of documents that need to be reviewed. But some critics contend that keyword searches may fail to identify potentially relevant information. As Magistrate Judge Paul Grimm observed in his 2008 *Victor Stanley* decision, "while it is universally acknowledged that keyword searches are useful tools for search and retrieval of ESI, … there is a growing body of literature that highlights the risks associated with conducting an unreliable or inadequate keyword search or relying exclusively on such searches."

While no computer-assisted information retrieval (IR) system yet developed can simply scan through a mountain of data and infallibly identify exactly those documents that an attorney would deem relevant, many IR options exist beyond traditional keyword searches that can reduce the risks associated with using search terms. For example, established search algorithms, commonly called "Boolean" or "set-theoretic" models, make binary decisions regarding the responsiveness of documents based on various simple tests, such as the presence of keywords within a certain distance of one another, linked by AND, OR, and NOT. A document is judged as either responsive or not, with no middle ground. Other search approaches, often broadly gathered under the rubric of "concept searches," move beyond this paradigm in a variety of ways.

Mitigating the Risks Associated with Using Search Terms

There are several ways to mitigate the risks of using traditional keyword searching. While not all of these options are appropriate in every case, consideration should be given to the following factors:

- Even within the Boolean paradigm, search tools can take advantage of "fuzzy" text comparisons and auxiliary structures such as thesauri to expand upon the queries generated by attorneys and thereby deal with the misspellings, optical character recognition (OCR) errors, synonymy (multiple words for a single concept) and polysemy (multiple meanings for a single word) that plague keyword searches. The keyword search also can be enhanced by interviewing key custodians about

the language that they use in correspondence, and by consulting with electronic discovery experts who are trained in keyword search development.

- "Algebraic" IR methods generate a measure of how similar each document is to what a query ideally seeks, thereby enabling the tool to *rank* documents by relevance rather than simply assigning them to two undifferentiated camps: responsive and non-responsive.
- "Probabilistic" or "Bayesian" search algorithms make use of more user input than simply the initial query in order to estimate a particular document's relevance.
- Tools such as domain name restrictions, discussion threading, topic clustering, people analytics, and analytics to identify duplicate copies of files can facilitate effective review.

The advantages of choosing a search system that is suited to your particular problem can be significant. By ranking documents rather than simply tagging all responsive documents as equally good, algebraic and probabilistic algorithms facilitate faster identification of key documents. By taking into account reviewers' tagging decisions and not simply the initial query, these systems reduce the need for humans—who charge by the hour—to keep repeating their recommendations. (Systems can be calibrated to permit some redundancy, to ensure that tagging mistakes do not poison an entire search.) Seemingly abstruse discussions of IR algorithm improvements quickly resolve themselves into bottom-line impacts in terms of the cost and time required to respond to discovery requests.

The Risks Associated with Using Concept Searching

If the use of concept searching can reduce risks and costs associated with retrieving relevant documents, why are these tools not already more popular among lawyers? There are several remaining risks:

- There is not yet much case law certifying non-Boolean search methods as acceptable. While this is changing—for example, in a 2007 opinion, Judge Facciola of the District Court for the District of Columbia noted that "recent scholarship … argues that concept searching, as opposed to keyword searching, is more efficient and more likely to produce the most comprehensive results," —the use of concept searching remains untested in the law, and opposing lawyers may balk at its use. On the other hand, successfully challenging the thoughtful use of Bayesian concept clustering is much more difficult and complicated than simply pointing to omitted search terms in a Boolean search string.
- Boolean search tools are ubiquitous and fungible. By contrast, software packages supporting concept searching are less well known. Each package comes with its own idiosyncratic user interface (no universally agreed syntax here). And there are significant differences between mathematical concept searching and thesaurus based concept searching. Thus, choosing a quality vendor can be a bigger challenge.
- Boolean searches generate output that lawyers understand—or at least think they do. The output of a keyword search is a list of documents that match the search query and, more importantly, a list of documents that do not. When asked if all responsive documents have been produced, an attorney who trusts keyword searches implicitly will answer, without reservation, "yes." The *ranked* output from an algebraic or probabilistic search provides no bright lines and, thus, requires more nuanced communication with the court. This distinction just makes explicit the uncertainties that are already present in Boolean searches that the binary output obscures: user-defined queries are

far from perfect, and the statement that no document responsive to the *search query* has been withheld is a far cry from certification that no document responsive to the *document request* has been withheld. Nontraditional search algorithms do not create this uncomfortable truth; they just bring it to the fore.

Regardless of the information retrieval methodology selected, documentation of which model was chosen and how it was implemented is an important tool to facilitate defense of the chosen process.

So, which search method is right for you? That can vary based on the types and volume of documents searched, the time frame and budget permitted, your aversion to risk, and your organization's comfort with technology, among other factors. But if the universe to be searched is large and costs are likely to be scrutinized, consideration should be given to making use of concept search technology, especially as prices for such technology have fallen dramatically.

For inquiries related to this Tip of the Month, please contact Tom Lidbury at tlidbury@mayerbrown.com, Jason Fliegel at jfliegel@mayerbrown.com or Zach Ziliak at zziliak@mayerbrown.com.

Learn more about Mayer Brown's Electronic Discovery & Records Management practice or contact Anthony J. Diana at adiana@mayerbrown.com, Michael E. Lackey at mlackey@mayerbrown.com or Thomas A. Lidbury at tlidbury@mayerbrown.com.

Please visit us at www.mayerbrown.com