New York Law Journal
TECHNOLOGY TODAY
Tuesday, April 22, 2008
ALM

Assessing Alternative Search Methodologies

• ELECTRONIC DISCOVERY •

BY H. CHRISTOPHER BOEHNING AND DANIEL J. TOAL

It may be that alternative search methods eventually will surpass the performance of keyword and Boolean searches, but that day does not yet seem to have arrived. The independent research conducted to date suggests that, for the time being at least, nothing beats Boolean, particularly when used as part of an iterative process.

As the burdens of e-discovery continue to mount, the search for a technological solution has only intensified. The holy grail here is a search methodology that will enable litigants to identify potentially relevant electronic documents reliably and efficiently.

In an effort to achieve these often competing objectives, litigants most commonly search repositories of electronic data for documents containing any number of defined search terms (keyword searches) or search terms appearing in a specified relation to one another (Boolean searches). These search technologies have been in use for years, both in litigation and elsewhere, and accordingly are well understood and widely accepted by courts and practitioners.[1]

But keyword and Boolean searches are far from perfect solutions; they are blunt instruments. Such searches will identify only those electronic documents containing the precise terms specified. These methodologies therefore will not catch documents using words that are close, but not identical, to the specified search terms, such as abbreviations, synonyms, nicknames, initials and misspelled words.

On the other hand, using more search terms may reduce the risk that an electronic search will miss a relevant document, but only at the price of increasing—often quite dramatically—the number of irrelevant documents found in the search. This is a serious problem because counsel must manually review whatever documents the searches yield in order to sift out nonresponsive materials, make privilege determinations and designate confidential documents. Keyword and Boolean searches thus require a careful balance to be struck: Unduly restrictive searches may miss too many responsive documents while overbroad searches threaten stratospheric discovery costs.

H. Christopher Boehning *and* **Daniel J. Toal** *are litigation partners at Paul, Weiss, Rifkind, Wharton & Garrison LLP. Associate* **Jason D. Jones** *and* **Aaron Gardner,** *the firm's discovery process manager, assisted in the preparation of this article.*

Christopher Boehning

Daniel J. Toal

Against this backdrop, courts and litigants understandably have been intrigued by the claims of those promoting alternative search technologies, such as "concept searching." The vendors of such technologies suggest their search strategies are able to identify the overwhelming majority of responsive documents while virtually eliminating the need for lawyer involvement in the review process.

Such claims strike many in the legal community as too good to be true. And their skepticism is appropriately heightened because the precise methodologies that such vendors use often are shrouded in mystery, owing to their stated desire to safeguard their proprietary processes and techniques. But this also means their tantalizing claims cannot readily be subjected to independent scrutiny. The question thus posed—and still largely unexplored—is whether these alternative search technologies have anything to offer and, if so, how best to evaluate the competing technologies and the often sensational claims of their promoters.

To evaluate whether an alternative search technology might be helpfully employed in any particular case, it is first essential to understand how it works. Some of the principal alternative search technologies, which fall under the broad heading of "concept searching" methodologies, are as follows:[2]

• **Clustering.** Whereas keyword and Boolean searches mechanically apply certain logical rules to identify potentially relevant documents, clustering relies on statistical relationships, which results in documents containing similar words being clustered together in relevant categories. The clustering tool compares each document in a pool to "seed" documents, which have already been designated as relevant. The more words a document has in common with a seed document, the more likely it is to be about the same subject and therefore to be responsive.[3] Moreover, clustering tools generally rank documents based on their statistical similarity to the seed documents.

• **Taxonomies and ontologies.** A taxonomy tool is used to categorize documents containing words that are subsets of the topics relevant to a litigation. For example, if one of the topics of interest is "dogs," a taxonomy tool would capture documents that mention "golden retrievers," "poodles" and "chihuahuas." Ontology tools perform similar searches, but are not confined to identifying subset relationships. Building on the last example, an ontology tool would capture documents that mention "kennels" or "veterinarians."[4]

• **Bayesian Classifiers.** Bayesian search systems use probability theory to make educated inferences about the relevance of documents based on the system's prior experience in identifying relevant documents in the particular litigation.[5] The search results then would be ranked based on the predicted likelihood of their relevance to the litigation.

How Approaches Compare

These alternative search technologies may sound promising in concept, and the claims about their efficiency and accuracy likely add to their allure, but the question remains whether these approaches outperform the standard search approach.

Keyword searching (including with the use of Boolean connectors), its acknowledged limitations notwithstanding, has secured such widespread acceptance for a reason. As an initial matter, the technology and search methodology is well understood and familiar to anyone who has used Westlaw, Lexis or similar search engines. It therefore can be easily discussed with both opposing counsel and judges. The simplicity of keyword searching also doubtlessly promotes negotiated resolution of discovery disputes because the parties have less reason to fear that ignorance about the technology will lead them to strike a bad bargain.

But the simplicity of keyword searching is also its principal weakness. Keyword searches capture only documents containing the precise terms designated, which virtually assures that such a search will miss relevant documents. And, on the other side of the equation, keyword searches will mechanically capture every document—whether relevant or not—containing any search term. This means keyword searches may be both substantially under- and over-inclusive. Concept searching systems, by contrast, are not dependent on a particular term appearing in a document and therefore may locate documents a Boolean search would not. But they may suffer from other infirmities.

So how does concept searching stack up? The best evidence to date comes from the Text Retrieval Conference (TREC), which in 2006 designed an independent research project to compare the efficacy of various search methods.[6] In view of the prevalence of keyword and Boolean searches in litigation today, TREC was particularly interested in determining whether the alternative search methodologies outlined above were better than Boolean.[7]

As its starting point, the TREC study used a test set of 7 million documents that had been made available to the public pursuant to a Master Settlement Agreement between tobacco companies and several state attorneys general.[8] Attorneys assisting in the study then drafted five test complaints and 43 sample document requests (referred to as topics). The topic creator and a TREC coordinator then took on the roles of the requesting and responding counsel and negotiated over the form of a Boolean search to be run for each document request.[9]

In addition to the Boolean searches, computer scientists from academia and other institutions attempted to locate responsive documents for each topic utilizing 31 different automated search methodologies, including concept searching.[10] The results were striking. On average, across all the topics, the negotiated Boolean searches located 57 percent of the known relevant documents.[11] But none of the alternative search methodologies reliably performed any better. That is to say, for each topic, the Boolean search did about as well as the best alternative search methodology.[12]

Interestingly, although the Boolean searches generally outperformed the alternative search protocols, the methods did not necessarily retrieve the same responsive documents. In fact, when all of the responsive documents found by the 31 alternative runs were combined, TREC discovered that the alternative search runs collectively had located, on average, an additional 32 percent of the responsive documents in each topic.[13] As a result, while the Boolean search generally equaled or outperformed any of the individual alternative search methods, those searches also captured at least some responsive documents that the Boolean search had missed.

Cost Analysis

This suggests that even if alternative search methodologies have not yet been shown to beat Boolean searches, their use to supplement Boolean searches might increase the number of responsive documents located. But at what cost? The potential benefits of locating any additional documents through use of an alternative search methodology would still have to be weighed against the cost, both in money and resources, required to locate them.

The relevant cost here is not just the price of using the alternative search technology, but also the number of false positives identified by the approach (i.e., documents retrieved by the search, but turn out not to be responsive). Any automated search method—whether a keyword or concept search—will yield false positives, which counsel must review and filter out prior to production, which can be a costly process. It therefore is far from clear that use of an alternative search methodology in addition to a keyword or Boolean search will be appropriate in any particular case, a question the TREC study does not attempt to address.

For now, the available evidence suggests that keyword and Boolean searches remain the state-of-the-art and the most appropriate search technology for most cases. This seems particularly true when keyword or Boolean searches are used in an iterative manner, where litigants: (i) negotiate search terms and Boolean operators, (ii) run the agreed-upon searches, (iii) review the preliminary results, and (iv) adjust the searches through a series of meet-and-confers. This type of "virtuous cycle of iterative feedback" has been endorsed by courts and commentators alike.[14]

The intuition of the legal community that an iterative approach to electronic discovery promotes reliability and efficiency finds empirical support in the TREC study. As part of its study, TREC employed an expert tobacco document searcher who used an "interactive" search methodology.[15]

TREC found that the expert searcher located, on average, an additional 11 percent of the relevant documents beyond those that had been located by the initial Boolean searches, which means that an interactive Boolean approach ultimately located 68 percent of the relevant documents—far better than any of the alternative search methodologies.

Conclusion

It may be that alternative search methodologies eventually will surpass the performance of keyword and Boolean searches, but that day does not yet seem to have arrived.

The independent research conducted to date suggests that, for the time being at least, nothing beats Boolean, particularly when used as part of an iterative process.

That does not necessarily mean that alternative search technologies are not worth considering, either independently or along with Boolean or keyword searches. But practitioners would be well advised to carefully scrutinize the marketing claims of the purveyors of such technologies and to factor in often substantial direct and indirect costs of such approaches.

1. See, e.g., *Treppel v. Biovail Corp.*, 233 F.R.D. 363, 374-75 (S.D.N.Y. 2006).
2. See The Sedona Conference, The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery 217-23 (Aug. 2007).
3. Id. at 219.
4. Id. at 221-22.
5. Id. at 218-19.
6. Jason R. Baron, David D. Lewis, & Douglas W. Oard, TREC-2006 Legal Track Overview, http://trec.nist.gov/pubs/trec15/papers/LEGAL06.OVERVIEW.pdf (hereinafter TREC 2006 Overview).
7. Id. at 2.
8. Id. at 3.
9. Id. at 4-5.
10. Id. at 10-11.
11. Id. at 11.
12. Id. at 13. In the second year of the TREC project, it appears the negotiated Boolean search located an average of only 22 percent of the relevant documents per topic. And although the percentages have changed, it apparently remains the case that no alternative search run reliably outperformed Boolean. An overview of the TREC 2007 Legal Track should be made public shortly. See Jason R. Baron, EED Showcase: Discovery Overload, LAW TECH. NEWS (January 2008), available at http://www.commonscold.typepad.com/eddupdate/2008/01/edd-showcase-di.html.
13. Id. at 12.
14. George L. Paul & Jason R. Baron, Information Inflation: Can the Legal System Adapt?, 13 RICH. J. L. & TECH. 10, ¶¶50-56 (Spring 2007), available at http://www.law.richmond.edu/jolt/v13i3/article10.pdf. See, e.g., *Balboa Threadworks Inc. v. Stucky*, 2006 WL 763668, at *5 (D. Kan. March 24, 2006).
15. TREC 2006 Overview at 5, 11.