THE NATIONAL LAW JOURNAL
AUGUST 23, 2010
Online Feature
An ALM Publication

ISTOCKPHOTO/PLAINFIELD

What lawyers need to know about search tools

The alternatives to keyword searching include linguistic and mathematical models for concept searching.

BY MAURA R. GROSSMAN AND TERRY SWEENEY

It will come as no surprise to anyone who has handled complex litigation during the past five years that the volume of electronically stored information (ESI) that must be reviewed in the course of discovery can be staggering. It may be more surprising to learn that keyword search is not nearly as effective at identifying relevant information as many lawyers would like to believe. See David C. Blair and M.E. Maron, "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval Sys.," 28(3) Comm. of the ACM 289 (1985) (showing lawyers estimated their search had identified 75% of the relevant documents when only about 20% were found); Douglas W. Oard, et al., Overview of the TREC 2008 Legal Track (March 17, 2009), (showing Boolean search identified only 24% of the relevant documents); Stephen Tomlinson, et al., Overview of the 2007 TREC Legal Track (April 30, 2008), (showing Boolean search identified only 22% of the relevant documents).

Litigators today face severe challenges in identifying and producing documents responsive to requests for production, on time, within budget and without waiver of privilege. See, e.g., *In re Fannie Mae Sec. Litig.*, 552 F.3d 814 (D.D.C. 2009) (involving delayed production in which 400 search terms yielded 660,000 documents, costing $6 million — or 9% of annual budget — to review); *Mt. Hawley Ins. Co. v. Felman Prod.*, No. 3:09-CV-00481, 2010 WL 1990555 (S.D. W. Va. May 18, 2010) (finding waiver of privilege for inadvertent production of 377 privileged documents in 346-gigabyte production). To assist lawyers in these efforts, there are a dizzying array of vendors and search tools on the market, each claiming to offer the "silver bullet." For time-strapped lawyers who have little — if any — interest in technology, sorting through the options can be overwhelming. But the consequences of getting it wrong — and using a shovel when one really needs a crane — can be severe, in terms of cost and otherwise. See, e.g., *In re Fannie Mae Sec. Litig.*, 552 F.3d 814 (D.D.C. 2009) (upholding contempt citation for failure to comply with deadline in stipulated discovery order).

Are all search tools and methods created equal? Do they all achieve the same results? How can attorneys become sufficiently comfortable using search tools so they can certify that, "to the best of [their] knowledge... formed after a reasonable inquiry," their response to a document request is "complete and correct," and that they have produced everything — or as close to everything as possible — that is responsive to the request? Fed. R. Civ. P. 26(g)(1)(A).

The proper search technology, coupled with a sound process, can make a huge difference to the quality, cost and speed of production. But to leverage technology, match the right tool to the right problem and implement a defensible process that is likely to yield the optimal result, it is necessary to understand something about how different search methods work and their strengths and weaknesses.

TYPES OF SEARCH TOOLS

The most common search strategy employs keywords, whereby documents are searched against a list of words — or word combinations — generated on the basis of a production request. If any of the words (or, in some cases, their variations) are found in a document, the document is retrieved. Anyone who has used Google is comfortable with this approach, but also knows that the results are imperfect because many of the documents retrieved will be irrelevant. Boolean search expands on keyword search by retrieving documents using words in specific combinations. Keywords are connected by logical operators, such as "and," "or," "but not" or "within 'n' words of."

Unfortunately, keyword and Boolean searches have limitations. Because they are word-based, they often fail to identify responsive documents because the author used different words to discuss the subject. This can happen when different communities refer to the same subject differently;

research and development may use different language than sales and marketing to refer to the same product. There also may be deliberate attempts to disguise the subject of the communication — for example, by using code words. Attorneys may simply lack familiarity with all the ways the subject can be discussed and may fail to search for misspellings and abbreviations. On the opposite end of the spectrum, words that seem relevant to the request may actually have different meanings when used in other contexts and will therefore bring back "junk."

Considering these limitations, it is no wonder that searching for concepts — instead of words — has gained popularity. There are many varieties of concept search, each with pros and cons. Some of the most common are the following.

A VARIETY OF MODELS

A thesaurus can be incorporated into search software to automatically expand keywords to include synonyms. Taxonomies and ontologies are like thesauruses "on steroids." They reflect how words and concepts relate to each other, either through hierarchies of classes and subclasses or through nonlinear relationships. For example, a tool with a built-in taxonomy would find documents with the word "poodle" in a search for "dogs." A tool with an ontology would identify documents about "bats" in a search for "baseball." Linguistic models rely on how words are used in documents to determine their context. While they can compensate for variability in word usage, they can also yield high volumes of "false hits," and therefore their effectiveness has been limited.

Developed to solve information retrieval and content classification challenges outside the legal context, mathematical and statistical approaches have recently been refined and applied to the problem of search in e-discovery. Their common characteristic is that rather than using language to determine related content and context, they rely on complex mathematical analysis.

Bayesian classifiers use statistical probability models that learn by analyzing document content based on such factors as the location, frequency and proximity of words. They use this information to compute a mathematical "thumbprint" of the concepts contained in the documents.

Latent semantic indexing uses a mathematical technique known as principal component analysis to identify groups of words and word combinations with similar meaning. The outcome is like using a thesaurus and translation dictionary to identify similar terms and phrases, but the process is fully automatic and language-independent.

Mathematical models such as these are useful for aggregating like documents, which can help reviewers make faster and more accurate judgments about responsiveness, privilege or other attributes. Search tools using statistical models can also be used effectively to "find more documents like this," when provided with specific document exemplars.

Variants on this approach, referred to as "machine learning tools," use "seed sets" of documents previously identified as responsive or unresponsive to rank the remaining documents from most to least likely to be relevant, or to classify the documents as responsive or nonresponsive.

A few things to bear in mind about concept search tools: They may not always perform as expected on small volumes of data or on data that must be collected and reviewed on a rolling basis. Moreover, when documents related to important issues are underrepresented in the document collection, concept search may not be the best approach for finding them. This is why keyword or Boolean search can still occupy a place in locating the "smoking gun," if attorneys know what they are looking for.

Concept search tools can be effectively combined with information obtained from metadata — such as custodian, author and date sent — to reveal relationships within the document collection. It is possible to graphically illustrate which custodians have had contact with which other custodians to discuss important topics within particular time frames. These tools can be especially useful in the early stages of a case or investigation to reveal unknown relationships, or to test case theories against the facts.

THE TREC LEGAL TRACK

Since 2006, the Legal Track of the Text Retrieval Conference, administered by the U.S. National Institute of Standards and Technology, has sought to study the application of information retrieval methods to e-discovery. The Legal Track brings together lawyers, e-discovery vendors and researchers from around the world to objectively evaluate different information retrieval methods. The results of the 2009 exercise are promising; they show that it is possible to achieve a high level of accuracy in identifying documents responsive to requests for production using automated and semi-automated methods. See Bruce Hedin, et al., Overview of the TREC 2009 Legal Track (July 19, 2010).

In a June 30 open letter to law firms and companies in the legal technology sector, the Sedona Conference observed that, "[f]or e-discovery service providers, law firms and corporate counsel, participation in the TREC Legal Track offers an unprecedented opportunity to be at the forefront of an important movement to evaluate document review processes, create industry best practice standards and, in so doing, provide the legal community...reliable information in the emerging field of large-scale document review and electronic discovery." http://trec-legal.umiacs.umd.edu/TREC_2010_Open_Invitation.pdf.

The key to success in search includes using the proper tools and methods for what one is trying to accomplish. Selection and application of search technology can require expertise, and attorneys should not hesitate to request it. Although choosing the appropriate tool is necessary, it is not sufficient; one also needs to implement a sound process. There is simply no substitute for careful planning, informed legal judgment and appropriate quality control, especially when timelines and budgets are tight and stakes are high.

Recent case law reminds us that the goal of e-discovery is not perfection, but reasonableness and proportionality. See *Pension Comm. of Univ. of Montreal Pension Plan v. Banc of Am. Sec.*, 685 F. Supp. 2d 456, 461 (S.D.N.Y. 2010); *Rimkus Consulting Group v. Cammarata*, 688 F. Supp. 2d 598, 613 (S.D. Texas 2010). Not every matter will warrant a "Cadillac" approach, but when properly applied, the right search technology can assist with the heavy lifting.

Maura R. Grossman is counsel practicing at New York's Wachtell, Lipton, Rosen & Katz and a co-coordinator of the 2010 TREC Legal Track. Terry Sweeney is the InfoDox platform manager at IE Discovery Inc. and has more than 25 years of experience in the legal technology services industry. The views expressed are solely the authors'.