Pros And Cons Of Computer-Assisted Review

Law360, New York (June 01, 2012, 12:02 PM ET) -- Following Magistrate Judge Andrew Peck's decision in Da Silva Moore v. Publicis Groupe approving the parties' agreement to the defendants' use of computer-assisted document review, which just recently was affirmed by the district court, the legal press and blogosphere have pronounced dead the decades-old practice of keyword searching and exhaustive manual review and announced the dawn of a new era of automated computer-aided review, which — we are told — is faster, cheaper and more accurate.

But any accuracy improvements and cost reductions depend on lawyers' ability to choose the appropriate tool for a particular case. Articles and blogs throw around terms like "predictive coding," "seed sets" and "statistical sampling" without clear explanation. Choosing the best tool for a particular case — and using it effectively — requires a deeper understanding of how these technologies stack up, both against each other and against more traditional manual review. And most importantly, it requires counsel to have experience leveraging these sophisticated tools in a defensible and cost-effective manner.

A look under the hood of currently available computer-assisted review platforms reveals that they employ one of three general approaches: (1) "rules-based" searching — think keywords on steroids; (2) "concept-based" searching; and (3) statistical active learning. The term "predictive coding," which has been used to describe some or all of these approaches, and other buzzwords provide little insight into the factors that a lawyer should understand when deciding which technology to use and how to use it.

Rules-Based Approach

Rules-based tools, such as the platform offered by Valora Technologies, code documents using a complex tree of nested "if-then" statements. A simple example: "If the document is an email from Jennifer In-House Counsel to an @outsidefirm.com email address, then code as privileged."

These rules are created through an iterative process by the vendor's staff of attorneys and technicians working together with the legal team, who manually review the system's coding and provide feedback to the vendor. This process of manual review and refining the rules is repeated until the automatic coding of the sample matches the legal team's coding choices. The rules are then used to code the entire document review population. As a final check, the legal team reviews a sample of the documents that were coded as not relevant, and the rules are adjusted to account for any relevant documents located in this sample.

The flexibility and nuance that can be incorporated into the rules enhances the efficiency and defensibility of this approach. Indeed, the rules can be easily tweaked to account for changes in scope as a matter progresses, due to, for example, an unanticipated outcome in a discovery dispute or the receipt of additional document requests.

The rules-based approach can be significantly less costly for voluminous reviews because it can effectively replace first-level review typically performed by contract attorneys. But depending on the number of relevant issues and their complexity, the iterative process of developing the rules can be time-intensive.

Moreover, from a defensibility perspective, an opposing party or regulator may demand production of the rules themselves or insist on being involved in the process of creating the rules, just as keyword lists often are produced to or developed with input from the opposition. In fact, several vendors of rules-based platforms encourage production of the rules and provide expert witnesses who can explain them in court.

Concept-Based Approach

Concept-based review platforms, such as Relativity Assisted Review, start by crunching the data to automatically create what is known as a “latent semantic index,” which essentially indexes the documents based on groups of symbols and patterns that appear in each document. In the same way that a book’s index cross-references topics to one another, the latent semantic index links documents that have similar concepts.

For example, the index could link documents with the word “Raptor” to documents with the word “Lucy” (names of Enron subsidiaries), even if the words never appear in the same document, because similar phrases such as “shell” or “balance sheet” appear in relation to each of those words in the respective documents in which they appear. This is a simple example, but in practice, the concept-based latent semantic index is highly sophisticated and effective at identifying documents with similar concepts as one another.

The legal team manually reviews and codes a “seed set” of documents. After the manual review, the system codes the remaining documents by extrapolating the reviewers’ coding choices to documents that are regarded as similar by the latent semantic index. Then, a new sample of documents is randomly selected from the remaining documents.

The legal team reviews the new sample, after which the system again automatically codes the remaining documents, and the process repeats. When the manual review of the sample of remaining documents yields sufficiently few responsive documents, the system is considered stable, no more iterations are necessary, and all of the documents have been coded even though only a relatively small sample has been manually reviewed.

Concept-based tools offer tremendous efficiency gains when compared with traditional keyword searching and exhaustive manual review. The concept-based tools can identify documents that might have been missed with keyword searching or a rules-based approach because the latent semantic index can identify relationships previously unknown to the legal team or the client.

And the entire process entails manual review of only a portion of the document set while yielding results that are more accurate and precise than traditional linear review. However, the process still is time-intensive — a dozen or more iterations can be required, and the number of iterations is highly dependent on the types of documents in the data set.

In addition, production of the entire seed set — regardless of whether the documents were coded responsive — may be necessary to give sufficient comfort to opposing counsel, and the court, about the accuracy and reliability of the review. Judge Peck's opinion in Da Silva Moore praised the defendants' willingness to share the seed set and the results of defendants' manual coding of the seed set.

The court explained that "transparency allows the opposing counsel (and the court) to be more comfortable with computer-assisted review, reducing fears about the so-called 'black box' of the technology" and recommended that "counsel in future cases be willing to at least discuss, if not agree to, such transparency."

Statistical Active Learning

There are two parts to the functionality of statistical active learning tools, such as the products offered by Equivio or Backstop. First, these tools "learn" from the legal team's coding of sampled documents, and second, they apply statistical techniques, similar to those that credit rating agencies use to determine personal credit scores, to group the documents by probable relevance. The Terminator movies involved computer systems that learned so well that they no longer needed humans to function. By contrast, active learning tools depend on the coding decisions of the legal team to learn to distinguish documents that are relevant from those that are not (and they won't travel back in time to kill you).

An initial set of key terms and documents are loaded into the system, which uses that information to initially score the entire document population based on as many as 20,000 properties of each document. Just as a credit score is based on the probability that a person will make future payments on time, statistical active learning systems assign a "relevance score" to each document, say from zero to 100, based on the probability that the document is relevant.

The system then selects a sample of documents for the legal team to manually review. Unlike with concept-based tools, however, the sampling is not random. Rather, the system selects the documents in the sample to "test" how well it scored the documents. As the attorneys manually code the sampled documents, the system "learns" by graphing a "hyperplane" that correlates the values of the 20,000 properties of each document with the attorneys' coding decisions. Those ubiquitous x-y graphs in Algebra II were in two-dimensional space; by contrast, the hyperplane is a graph in 20,000-dimensional space.

After the legal team manually reviews the sample, the system re-scores the entire population of documents based on their "distance" from the hyperplane in 20,000-dimensional space, and then selects a new sample of documents that the legal team reviews to test the system's scoring. The process is repeated until the system reaches statistical stability, which occurs when the attorneys' coding of the sampled documents is sufficiently consistent with the relevance scores assigned by the system.

Once the system is statistically stable, the documents are grouped by score range to identify the probability that the documents in each group are relevant, which enables the legal team to focus its review to quickly find key documents without wasting time and money reviewing thousands of irrelevant documents. But although these powerful statistical techniques can provide useful insights into the document population in the aggregate (e.g., that 96 percent of all relevant documents are in some subset of the review set), they cannot definitively code each document, just as credit scoring enables lenders to determine that, say, 95 percent of people with FICO scores above a certain threshold will pay their mortgage on time, but cannot provide any insight into whether a particular individual will do so.

Recent articles in the legal press have conflated concept-based searching with statistical active learning. This confusion may stem from the fact that the Recommind Axcelerate platform that the parties proposed to use in Da Silva Moore is a hybrid that combines active learning functionality with the workflow of concept-based tools to definitively code each document rather than group the documents by probable relevance.

Recommind's platform iteratively selects samples of documents for manual review that the system determines are relevant based on what it learned from the legal team's coding of prior document samples. After each sample is manually reviewed, the system revises its coding of the remaining population and selects a new sample. This sampling and manual review process is repeated until the system determines that there are no more relevant documents in the population. The end result is that the legal team has manually reviewed and coded all of the relevant documents but only a small fraction of nonrelevant documents.

Both the Recommind platform and concept-based tools definitively code each document, but often entail manual review of a substantially larger percentage of the documents (25 percent or more) than is required for statistical active learning tools (15 percent or even less). In addition, statistical active learning tools add value by providing the legal team with an understanding of the entire document population early in the lifecycle of a case, which can yield significant cost savings and review benefits. For example, the legal team might decide to review only the highest scored documents that are most likely to be significant while lower scored documents are reviewed by contract attorneys or not at all.

Although statistical active learning tools may get up and running more quickly, both concept-based tools and Recommind's technology can more easily be adjusted mid-stream if new documents are collected and added to the review set or if the scope of the review changes as a result of, say, the receipt of new document requests or the introduction of a new issue into a case. By contrast, statistical active learning tools must be re-taught from scratch because the statistical inferences drawn from samples are no longer valid if the population or scope of review substantially change.

Statistical active learning tools raise similar defensibility issues as concept-based tools. Counsel may need to share the seed set with opposing counsel and the court to provide the necessary transparency and comfort with the process. And in weighing the costs and benefits of any computer-assisted review technology, counsel must also consider a host of other issues, including whether the vendor requires counsel to also use that company's contract attorneys or document review platform.

Acceptance of computer-assisted review is growing, as is the recognition of its value and effectiveness. Four months before Judge Peck's decision in Da Silva Moore, the U.S. District Court for the Southern District of New York implemented a pilot program designed to reduce the high costs and delays of litigating complex civil cases.

The program incorporated computer-assisted review, expressly identifying "concept search," "machine learning" and "other advanced analytical tools" among the approaches for the "search and review of electronically stored information" that parties in complex civil litigation may consider in connection with a Rule 26(b) conference. And in January 2012, the Federal Trade Commission proposed revisions to its Rules of Practice to permit the use of "concept searches, predictive coding, and other advanced analytics."

Not everyone is welcoming this development with open arms. The plaintiff in Da Silva Moore sought Judge Peck's recusal because of his well-known views on computer-assisted review, but the district court recently denied the motion.

Plaintiffs in a large putative class action before Magistrate Judge Nan Nolan in the Northern District of Illinois — Kleen Products LLC v. Packaging Corp. of America — have sought to require the defendants to use computer-assisted review instead of more traditional keyword searching techniques. And on April 23, 2012, a Virginia state court — in Global Aerospace Inc. v. Landow Aviation LP — granted the defendants' motion requesting permission to use computer-assisted review over the plaintiffs' objections.

Despite resistance from some quarters, there can be little doubt that computer-assisted review is the future of e-discovery. In the absence of authoritative best practices, courts and users of computer-assisted review will continue to grapple with how best to validate the results of the various types of tools. Lawyers who understand how to use these technologies to deliver defensible and cost-effective approaches that are well-tailored for each case will have an advantage as e-discovery continues to play a critical role in modern litigation.

--By David L. Breau, Sidley Austin LLP, and Adrian Fontecilla, Crowell & Moring LLP

David Breau is an associate in Sidley Austin's New York office and a member of the firm's e-discovery task force. Adrian Fontecilla is an associate with Crowell & Moring in the firm's Washington, D.C., office and a member of the Seventh Circuit Electronic Discovery Pilot Program Committee.

The opinions expressed are those of the authors and do not necessarily reflect the views of the firm, its clients, or Portfolio Media Inc., or any of its or their respective affiliates. This article is for general information purposes and is not intended to be and should not be taken as legal advice.