Predictive Coding Helps Companies Reduce Discovery Costs

Recent Court Decisions Open Door to Wider Use by Businesses to Cut Costs in Document Discovery

By John Tredennick

As companies struggle to manage exploding volumes of electronic content, legal and IS departments have had to find new ways to comply with data discovery obligations. Increasingly, global litigation and regulatory investigations require the production of a range of company documents and data. Courts have stepped up requirements to preserve and share these files, imposing multi-million dollar sanctions for noncompliance.

This has spawned an enormous industry built around electronic discovery. Companies spend billions on collecting, organizing and reviewing their data. Fully 73 percent of that spend is on reviewing documents to find those relevant to the matter and protect those that might be subject to the attorney-client privilege, RAND recently found.

Attorneys have long felt it necessary to manually review each document being considered for production. When data volumes were still relatively small, this expense was a rounding error on much bigger legal budgets. However, as document volumes increased into the millions, attorney review costs mushroomed to become the largest part of the discovery spend.

New Approaches to Reducing Review Costs

Faced with increasing volumes, corporate counsel began looking for new ways to cut review burdens. They began using search techniques designed to better target relevant documents and exclude junk. Then they employed clustering to help group documents with similar content. Some sent review work offshore. All the while, volumes kept increasing, outpacing counsel’s ability to keep up.

Recently, a more-promising approach to reduce document volumes has emerged. Called variously predictive coding, computer-assisted review and technology-assisted review, the technology has been proven to streamline the
review process dramatically, reducing both the quantity of data involved and the
time required to sift through it.

As promising as this technology is, however, lawyers and their clients have been
slow to adopt it. Their reluctance is often attributed to risk management or, in
lawyer talk, what is called “defensibility.” Lawyers fear that, if called on in court to
defend their data discovery, predictive coding would be too untested a
technology to hold up.

Seemingly overnight, that mindset changed. On Feb. 24, 2012, a federal
magistrate-judge in New York City, Andrew J. Peck, issued the first judicial
opinion anywhere to endorse the use of predictive coding in electronic discovery.
It was the shot heard ‘round the litigation world, prompting headlines declaring
the ruling a breakthrough and the judge a trailblazer.

As Judge Peck himself had observed just two months earlier in a legal trade
magazine, "While anecdotally it appears that some lawyers are using predictive
coding technology, it also appears that many lawyers (and their clients) are
waiting for a judicial decision approving of computer-assisted review."

Without doubt, Judge Peck ushered in the next generation of e-discovery search
and review. In the months since his decision, the business world has shown a
heightened interest in predictive coding. Other judges have picked up the mantle
and considered its use in their cases. Industry vendors have scrambled to add
predictive coding to their product rosters.

As important as was the ruling’s outcome, so too was its reasoning. For business
leaders, IT professionals and in-house lawyers, the opinion provides a primer on
the reasons for using predictive coding and the processes underlying it.

How Big Data is Changing Litigation

To understand the importance of predictive coding, it helps to step back and
consider how big data is changing the nature of litigation.

Before a trial, court rules allow litigants to “discover” each other’s evidence.
Litigants must identify their potentially relevant documents and hand them over to
their opponents. They can withhold documents protected by attorney-client
privilege. As for all the documents that are neither relevant nor privileged, they
can be ignored.

For decades, document review was done manually, with attorneys eyeballing
each document and deciding whether it should be produced. But as paper turned
into data and data grew into big data, manual review was no longer feasible. Today, major corporate lawsuits routinely involve gigabytes of data and sometimes even terabytes.

To help sift through it all, attorneys turned to technology. Their objective, as Judge Peck observed, was “to identify as many relevant documents as possible while reviewing as few non-relevant documents as possible.”

Attorneys most commonly use keyword searching. But keyword searches are imprecise. The way lawyers choose keywords, Judge Peck wrote, “is the equivalent of the child’s game of ‘Go Fish.’” Further, keywords are often over-inclusive, he noted, finding relevant documents but also large numbers of irrelevant ones.

Judge Peck cited a 1985 study by scholars David Blair and M.E. Maron. Experienced searchers were instructed to use keywords to retrieve at least 75 percent of relevant documents from a collection of 40,000. Although the searchers believed they had succeeded, their actual recall was just 20 percent.

By comparison, predictive coding is more effective and less expensive, Judge Peck said. In fact, studies have shown predictive coding to be at least as accurate as human review, if not more accurate, he noted.

“Computer-assisted review appears to be better than the available alternatives, and thus should be used in appropriate cases,” Judge Peck concluded. While computer-assisted review is not perfect, he added, court rules do not require perfection. The overarching goal is to “secure the just, speedy, and inexpensive determination” of lawsuits.

**How Predictive Coding Works**

Unlike manual review, which is mostly done by junior staff, computer-assisted coding involves senior members of the litigation team who review and code a seed set of documents. Software uses that seed set to code other documents. The process is iterative, with the lawyers providing feedback on the computer’s coding and the computer then coding additional documents.

The case decided by Judge Peck, *Da Silva Moore v. Publicis Groupe*, illustrates the process. To review a collection of some 3 million emails, the parties proposed a predictive-coding protocol. Judge Peck approved the protocol and included it with his written ruling. While there is no single right way to do computer-assisted review, the protocol provides insight into how the process can work.
To begin, the attorneys would identify a small number of emails to serve as seed sets representative of the categories to be reviewed and coded. The seed sets – one per issue – would then be used to begin training the software. The software uses each seed set to identify and prioritize all similar documents within the collection. The lawyers would then review at least 500 of the computer-selected documents to further calibrate the system.

Once the system is trained, it would proceed to identify relevant documents. Documents it identifies would be reviewed manually before they are produced to the opposing side. As a final quality-control step, the accuracy of the process would be verified using both judgmental and statistical sampling.

**Training the Software and Checking Results**

Predictive coding is an iterative process. After the software identifies the first set of potentially relevant documents, the parties review the results. The software then analyzes the new tagging and finds a second set for testing, then a third and a fourth.

Here, the protocol suggested that the process be repeated seven times. The key was to watch the change in the number of relevant documents predicted by the system after each round of testing. Once that number dropped below a delta of 5 percent, the parties could stop. That would indicate the system was stable, with further review unlikely to bear much fruit.

At that point, the process would move from computer-assisted to human-powered review. The producing party would review all of the potentially responsive documents and produce accordingly.

As a final check, the parties would focus on the documents the system said to ignore. Of those, they would check a random sample (2,399 again) to see how many were, in fact, responsive.

**What this Means to Business Leaders**

Predictive coding techniques can reduce document populations dramatically, often by more than 50 percent. This translates into substantial savings on review costs, easily shaving millions off total annual corporate e-discovery costs. It is not ideal for every case – it particularly shines in the larger ones – but it should be considered by every corporation facing litigation and regulatory inquiries.
The process outlined in Judge Peck’s ruling is not the bible of predictive coding. There are a variety of viable approaches. However, law is a profession built on precedent. First opinions carry significant weight.

For business leaders, the significance of the case is its seal of approval for predictive coding technology. That should lead to its wider acceptance and use. With data volumes increasing dramatically, predictive coding is the best weapon businesses have to cut the cost and time of review. In their battle to conquer big data, that is a significant victory.

John Tredennick is the founder and CEO of Catalyst Repository Systems, an international provider of multi-lingual document repositories and technology for electronic discovery and complex litigation. Formerly a nationally known trial lawyer, he was editor-in-chief of the best-selling book, “Winning with Computers: Trial Practice in the Twenty-First Century.” Recently, he was named to the Fastcase 50 as one of the legal profession’s “smartest, most courageous innovators.”

-END-