New York Law Journal
A NEW YORK LAW JOURNAL SPECIAL SECTION
An ALM Publication
E-Discovery
WWW.NYLJ.COM
MONDAY, MARCH 16, 2015

When Considering TAR, It's Never Too Late

Even midway through discovery, time and money can be saved.

BIGSTOCK

BY STEVEN M. AMUNDSON AND MARK NOEL

STEVEN M. AMUNDSON *is a partner with Frommer Lawrence & Haug.* MARK NOEL, *a former IP litigator, is managing director of professional services at Catalyst.*

Technology-assisted review (sometimes referred to as "predictive coding") is earning a well-deserved reputation for its ability to reduce the time and cost of e-discovery review in complex legal matters. What many lawyers may not realize, however, is that it is almost never too late for TAR. This is especially true of the newer "TAR 2.0" systems that are faster, more flexible and practical for a wide range of cases. Even when started midway through a relatively small, technical review, TAR's impact can still be dramatic.

We saw this firsthand in a recent case in which we represented a generic pharmaceutical manufacturer that had been sued for patent infringement by a major

brand-name pharmaceutical company. The plaintiff claimed that our client's generic products infringed its patents.

Our firm had considered TAR in other cases, and recommended its use here. Only after manually reviewing nearly half the collection, however, did the firm receive approval to proceed with TAR. Even that late in the game, TAR produced substantial savings in time and cost.

Starting With Linear Review

In this case, the total document collection to be reviewed (after applying agreed-to search terms and culling) numbered about 40,800 records. While this was not a huge collection by the standards of some cases, it was nevertheless a lot of documents to get through and would be a significant expense for our client.

Believing that TAR would enable us to get through the review more quickly and therefore at less cost, we recommended it to our client. But looming deadlines demanded that we start work on the document collection while our client considered our recommendation.

So without the benefit of TAR, we launched into a manual, linear review of the documents. Just before we reached the midway point in our review, we received approval to begin using TAR.

At that point, we had already reviewed some 18,200 of the 40,800 total documents. Had we started using TAR at the outset of the review, we might have hoped to avoid reviewing even that many documents in total. However, those 18,200 documents gave us the advantage of providing a ready-made set of seed documents to use to train the TAR algorithm.

Running the TAR Algorithm

After training the system, we ran the TAR algorithm against the remaining documents in the collection. We used Insight Predict, the TAR platform developed by Catalyst, which has the ability to use any and all previously coded documents (called "judgmental seeds" in TAR parlance) to start the process. Predict uses Continuous Active Learning (CAL), a machine learning protocol that has been shown in recent independent studies to consistently outperform older, more common TAR protocols. See, e.g., Gordon V. Cormack and Maura R. Grossman, "Evaluation of Machine Learning Protocols for Technology-Assisted Review in Electronic Discovery, in Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval" (July 2014).

Switching review methods in the middle of a case is not always something that can be done unilaterally. A party's ability to do that **may be limited** by the case management order or some other circumstances.

In practice, this means that there are not separate workflow phases for training the TAR system and for review. All documents that already have attorney decisions on them are fed into the system at the start, and the entire population is analyzed and ranked—a process that runs in the background and takes about seven minutes for a million documents.

After that first ranking is complete, the system provides a small batch of documents to reviewers that mostly contains the next-best, unreviewed documents the system can find, but also includes a few "contextually diverse" documents to make sure there are no topics or concepts left in the collection that go unexplored by reviewers. As reviewers complete their small batches of documents, the system continuously re-ranks the entire population in the background, incorporating those new coding calls to "get smarter" and improve its predictions.

Each time reviewers click a button for more documents, the system creates a new batch based on the most recently completed re-ranking. This means that the ranking is constantly improving and never stops learning. But from the reviewing attorneys' point of view, all they have to do is ask for more documents and then review batches that have a much higher proportion of relevant documents than they otherwise would have seen.

We proceeded along this track until we started seeing batches with few, if any, relevant documents. This is one of the indications that there are few relevant documents that remain unreviewed, and that the point of diminishing returns has been passed. Catalyst then helped us test our results by sampling the documents we had not reviewed. The statistical analysis of the sample review showed that we had achieved a very high "recall"—the review metric that describes how close we came to finding everything. Achieving such a high recall means that we found the vast majority of the relevant documents.

By the end of the TAR process, we had reviewed another 6,800 documents, beyond the 18,200 we had reviewed before beginning TAR. That meant that there remained another 15,800 documents that we never had to review. Put another way, once we fired up the TAR system, we only had to review 30 percent of the remaining documents before we were done. It saved 70 percent of the remaining expense and time the review would have otherwise required.

Again, by the standards of some large cases, that raw number may not sound like a huge savings. But a 70 percent savings on even a portion of a larger review quickly gets into some seriously large numbers. And even when you consider the savings to our client, we achieved a significant result.

Let's assume that a reviewer can typically get through about 50 records per hour. (Remember that this is a patent litigation, and technical documents typically take longer to review.) That would mean it would take 316 hours to review 15,800 records. Let's further assume that the average blended rate for the reviewers in this case was $250 an hour. At $250 an hour, that is a cost of $79,000. Here, the fees to the vendor, Catalyst, were about $10,000. Thus, the entire net savings was about $69,000.

The Law on Switching Horses

Of course, switching review methods in the middle of a case is not always something that can be done unilaterally. A party's ability to do that may be limited by the case management order or some other circumstances.

A leading case on this point is *Bridgestone Americas v. International Business Machines*, No. 3:13-1196 (M.D. Tenn. July 22, 2014). There, after initially screening a collection of over two million documents using key words, Bridgestone sought leave of court to use TAR for the remainder of its responsiveness review. IBM objected that this would be an unwarranted change in the original case management order and that it would be unfair to use TAR after doing the initial screening with key words.

After considering the parties' arguments, U.S. Magistrate Judge Joe B. Brown issued an order permitting Bridgestone to use TAR. Acknowledging that he was "allowing Plaintiff to switch horses in midstream," he reasoned that "openness and transparency in what Plaintiff is doing will be of critical importance." He noted that Bridgestone had agreed to provide its seed documents and that IBM is a "sophisticated user of advanced methods for integrating and reviewing large amounts of data."

"In the final analysis, the use of predictive coding is a judgment call, hopefully keeping in mind the exhortation of Rule 26 that discovery be tailored by the court to be as efficient and cost-effective as possible," Brown wrote. "In this case, we are talking about millions of documents to be reviewed with costs likewise in the millions. There is no single, simple, correct solution possible under these circumstances."

In our case, opposing counsel did not object to our initial manual, linear review followed by the switch to TAR. We had entered into a case-management protocol that required cooperation and collaboration in identifying initial search terms but specified nothing about what would happen after running the search terms against the agreed-upon custodians. Opposing counsel was well informed about TAR and used TAR for review of their client's documents.

The Specifics

Earlier, we outlined the general process we used in our review. For those of you who might be considering TAR, allow us to provide further detail on the workflow.

As noted above, the parties had agreed at the outset on certain search terms and custodians. Thus, we began by running the agreed-upon search terms against the files for the agreed-upon custodians and non-custodial data sources. After deNISTing and deduplication, that process yielded about 40,800 records for review.

Our reviewers then reviewed the records linearly, starting with the custodians and non-custodial data sources that we believed most likely to have a higher percentage of responsive records. Our reviewers made coding determinations ("responsive," "non-responsive" or "privileged") for approximately 18,200 records during the linear review.

It was at that point that we decided to employ TAR and, in particular, the Insight Predict TAR tool and document-relationship engine from Catalyst.

To train its search engine, Catalyst used our coding determinations for the approximately 18,200 records we had considered during the linear review. Catalyst's Predict was then set to automatically create batches of 50 records each that the TAR engine predicted were most likely responsive to the opposing party's production requests.

After a reviewer made his or her coding determinations for all of the records in a 50-record batch, the TAR engine utilized those coding determinations to update the algorithm and to continuously re-rank the entire population in the background. (This is the process called Continuous Active Learning.) Predict would then use the most recent re-ranking to create new batches on demand for the reviewers. This review-and-updating process continued until our reviewers repeatedly encountered batches containing few, if any, responsive records.

We then conducted targeted searches through the unreviewed records in an effort to locate other potentially responsive records. Our reviewers made coding determinations on the results of those targeted searches. After that, we had about 15,800 unreviewed records.

We next sampled (at a confidence level of 98 percent and a confidence interval of ± 3 percent) the entire population of about 40,800 records to determine the percentage of responsive records. That sampling indicated an overall richness rate of 15.3 percent (including privileged materials).

Based on that sampling, the entire population should include approximately 6,240 responsive records (including privileged materials). At that point, however, we had already coded about 6,400 responsive records (including privileged materials). Thus, we had an estimated recall of 102 percent and a confidence interval of ± 3 percent.

Based on that estimated recall, we decided to discontinue the review because the burden of or expense for continuing the review and finding any additional responsive records outweighed the likely benefit of any such records.

Conclusion

The law and practice surrounding the use of TAR in e-discovery continue to evolve. This case was somewhat unusual in that our legal team did not start out with TAR. It was only after manually reviewing nearly half the documents that we decided to switch to using TAR.

Even so, by using TAR, the team was able to eliminate the need to manually review nearly 40 percent of all the documents. That resulted in substantial cost savings to our client and time savings to our litigation team.