Five Questions to Ask Your TAR Vendor About Continuous Active Learning

How to Ensure that Your Review Platform Uses the Most Effective Technology

Thomas Gricks, Esq.
Managing Director, Professional Services, Catalyst Repository Systems
In their July 2014 paper, Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery, Maura Grossman and Gordon Cormack reported the results of their controlled TAR study in which they compared the effectiveness of a Continuous Active Learning (CAL) protocol against two first-generation (TAR 1.0) protocols that use one-time training. Their study found the CAL protocol to be more effective—most times much more effective—than the TAR 1.0 protocols.

With that evidence establishing the advantages of CAL, e-discovery vendors began jumping on the bandwagon. All of a sudden, it seems that every e-discovery vendor claims to use CAL or somehow incorporate CAL into its TAR protocols.

Despite these claims, there remains a wide chasm between the TAR protocols available on the market today. Knowing what you are buying is important. Grossman and Cormack found CAL to be far superior to TAR 1.0 at minimizing review while simultaneously maximizing recall, i.e., the fraction of responsive documents found. They further found CAL to be particularly more effective than TAR 1.0 in dealing with low-richness collections.

As a TAR consumer, how can you determine whether a vendor that claims to use CAL actually does? Here are five basic questions (plus a bonus) you can ask your vendor to ensure that your review effectively employs CAL.

1. Does your TAR tool use a control set for training?

Control sets are the hallmark of TAR 1.0, but wholly inconsistent with the concept of CAL. In fact, the use of a control set for training can often impair and complicate the TAR process.

To train a TAR 1.0 tool, you typically start by generating a random sample to represent the entire document population to some statistical degree of certainty. That random sample—a small fraction of the entire collection—is considered the control set. As documents are reviewed and coded, the progress of training is measured against the control set, which is re-ranked after every training round. Once it appears that training is having little impact on the ranking of the control set, the tool is considered to be stabilized. The one-time training effort concludes and the tool is used to rank the entire collection for production review.
By comparison, a true CAL process uses no control set. Every review decision is used to train the tool, and the entire collection (not a small subset) is constantly re-ranked and monitored. Only when it appears that you have reached your goal vis-a-vis the entire collection—or that training is having no further impact on the ranking of the entire collection and responsive documents are no longer being returned for review—do review and training cease.

2. **Does training focus on marginally relevant documents or highly relevant documents?**

Training that focuses on marginally relevant documents will not optimize the use of CAL. In fact, TAR protocols that focus on marginally relevant documents are typically not CAL and are generally less effective.

The predominant objective of reviewing marginally relevant documents is to determine where best to draw the line between relevant and non-relevant documents. The ultimate goal is to train an algorithm, called a “classifier,” to make that distinction, so the presumptively relevant documents can be separately reviewed for production. Generally, TAR protocols that use a classifier to segregate documents neither rank the collection nor train continuously through the attainment of review objectives. Thus, they would not be considered CAL.

As illustrated by the chart above, the classifier approach has two drawbacks. First, no matter how well the line is drawn, some number
of relevant documents (either cats or dogs) will fall on the other side of the line and never be seen in the review set. Second, the tighter you try to draw that line, the more time and effort it takes before you can even begin to review documents.

Conversely, the objective of CAL is to continuously use reviewer judgments to improve training and rank the collection, and to use the improved ranking to present the reviewers with better documents. This process continues iteratively until the review targets are achieved for the collection as a whole. To implement this protocol effectively, training and review focus primarily and specifically on highly relevant documents. There is no wasted effort, and every relevant document is available for review.

3. How often do you rank the collection during review?

The essence of CAL is its ability to harness reviewer judgments to rank the collection and return the best documents to the reviewers as early as possible. Every time CAL ranks the collection, reviewer judgments are leveraged to improve review and, in turn, to improve the next ranking. The process is cyclical and the results exponential.

Studies prove the obvious—the more frequent the ranking, the better the results. This phenomenon is akin to compounding interest; the more frequently interest is compounded, the more rapidly the benefit accrues. With CAL, the more frequently the collection is ranked, the more rapidly reviewers can take advantage of their collective decisions to successively feed judgments back to the tool for further refinement.

Among vendors, a tremendous disparity exists in the frequency with which they rank the collection (or the control set, with TAR 1.0). Catalyst, for example, can and does rank millions of documents in minutes. Reviewer judgments are used to rank the entire collection several times every hour throughout the review and training process.

Most other vendors rank only the control set (not the collection) during review and training, and subsequently rank the entire collection only once. Even worse, their process of ranking the entire collection can typically take several hours to complete.

“Studies prove the obvious—the more frequent the ranking, the better the results. This phenomenon is akin to compounding interest; the more frequently interest is compounded, the more rapidly the benefit accrues.”
4. Is it necessary to have a Subject Matter Expert do the training?

Although TAR 1.0 requires an SME for training, CAL does not. In fact, since all review is training with CAL, training by an SME would be prohibitive. CAL frees the SME to focus on more productive tasks and leave the bulk training to the reviewers. This enables immediate review and eliminates the time and expense associated with training a TAR 1.0 tool.

With TAR 1.0, training is a one-time effort, the results are driven by comparison to a finite control set, and the process dictates exactly which documents will be reviewed for production. This creates an inherent need for the consistency of a single decision maker with the knowledge and authority to establish the scope of the eventual review.

This is not the case with CAL, where every review judgment trains the tool. Reviewers see the same documents they would have seen after training using a TAR 1.0 protocol (perhaps more), and presumably make the same decisions. Because the tool is continuously learning from the reviewers’ judgments, the universe of documents passed to the reviewers is constantly refined to elevate those most likely to be produced.

The upshot of eliminating SME review is savings—of both time and money. TAR 1.0 typically requires at least two weeks of SME effort before the review team can review a single document. Given the billing rate of senior attorneys most likely to serve as SMEs, that effort will cost tens of thousands of dollars. With CAL, review and training are coextensive and start immediately, with no sunk cost for training.

5. How easy is it to run simultaneous TAR projects?

Another benefit of CAL is the ability to run simultaneous TAR projects and generate useful results almost immediately, at any point during the review and with little to no additional setup. With TAR 1.0, the process is much more cumbersome. If your vendor does not allow you to easily and quickly implement simultaneous TAR projects, you are not using CAL to its fullest potential.

Since review is training with CAL, very little is required to run simultaneous TAR projects covering different issues. Simply identify
each of the pertinent issues, and code the documents for each issue during review. The tool will use each judgment, and generate and maintain a separate ranking for each issue. Once you attain the review objective for one TAR project, you can focus on the next project. The existing ranking will make every successive project more efficient.

CAL can incorporate new TAR projects at any point during review and quickly generate results. Review and coding for a new project can start as soon as a new issue is identified. From that point forward, every review decision can include a judgment on the new issue. By focusing specifically on the new TAR project, review and training will quickly improve the ranking and return the best documents for further review.

TAR 1.0 is too cumbersome to do this effectively. With TAR 1.0, every project requires a separately coded control set against which to evaluate training. This makes simultaneous projects, especially new projects arising during review, difficult to implement.

**BONUS QUESTION: What is your process for handling rolling collections?**

The reality of modern discovery is that all of the documents to be reviewed for production are rarely available at the same time. Instead, documents are collected piecemeal and arrive on a rolling basis. An added benefit of CAL is its ability to incorporate new documents into the collection at virtually any point in the review without sacrificing previous effort. If the vendor suggests that a rolling collection presents any impediment to seamless review, the vendor is not making an efficient or effective use of CAL.

Rolling collections are a problem for TAR 1.0 protocols because they rely on control sets for training. A control set is intended to represent the entire collection. Since newly added documents change the character of the collection, the initial control set is no longer representative. Every time documents are added to a collection, a new, revised or additional control set needs to be generated. Even worse, if new documents are added after training is completed and a review set generated, it may be necessary to completely retrain the tool in addition to preparing a new control set.

"An added benefit of CAL is its ability to incorporate new documents into the collection at virtually any point in the review without sacrificing previous effort."

© 2015 Catalyst Repository Systems. All Rights Reserved.
CAL is not subject to these limitations. As new documents are received, they are simply incorporated into the collection and integrated according to the current ranking. As those documents are reviewed and coded through the continuous-learning process, the ranking adjusts to reflect the new information. No effort is lost. Every previous judgment remains intact and every subsequent judgment further improves the ranking.

CAL provides significant advantages over other TAR protocols, in both efficiency and effectiveness. So how can you be sure that your vendor is actually equipping your review project with all of the benefits of an efficient and effective CAL protocol? Just ask five simple questions—and throw in the bonus for good measure.

About the Author

Thomas C. Gricks is a managing director at Catalyst and a licensed attorney in the Commonwealth of Pennsylvania. He has been an active general commercial litigator for the past 23 years. He has practiced before both federal and state courts across the United States, and is currently admitted to practice in Pennsylvania as well as a number of federal jurisdictions, including the United States Supreme Court.

For the past several years, Mr. Gricks has devoted a substantial portion of his practice to e-discovery, with a particular emphasis on TAR. He argued the Global Aerospace case before the Circuit Court in Loudoun County, Virginia, which is the first case to permit the use of TAR over the objection of the opposing party. He is a Member of the Sedona Conference Working Group One, and is a frequent speaker on e-discovery and TAR topics. He has presented at continuing legal education seminars across the country, including the Sedona Conference Institute and the Georgetown Advanced eDiscovery Institute. He has also published a number of articles on e-discovery topics, including co-authoring The Implications of Rule 26(g) on the Use of Technology-Assisted Review, 7 FED. CTS. L. REV. 239 (2013).