Continuous Active Learning for TAR

The vast array of technology-assisted review (TAR) tools available in the marketplace, along with their associated jargon, can seem daunting to counsel. But using TAR in litigation and regulatory matters need not be. By implementing a continuous active learning (CAL) protocol, with a TAR tool that uses a state-of-the-art machine-learning algorithm, responding parties can quickly and easily identify substantially all of the relevant documents in a collection and minimize the potential for disputes with requesting parties.

MAURA R. GROSSMAN
OF COUNSEL
WACHTELL, LIPTON, ROSEN & KATZ

Maura focuses her practice on legal, technical, and strategic issues involving domestic and international e-discovery and information governance. She has served as a court-appointed special master, neutral, and expert on search-related issues, and her legal and scientific publications on TAR have been cited as authorities by federal, state, and international courts. Maura also teaches e-discovery courses at Columbia Law School and Georgetown University Law Center, and is a coordinator of the Total Recall Track at the Text REtrieval Conference (TREC).

GORDON V. CORMACK
PROFESSOR
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE,
UNIVERSITY OF WATERLOO

Gordon is a professor of computer science and serves as an independent e-discovery consultant. His primary research focus is on high-stakes information retrieval. Gordon is the co-author of Information Retrieval: Implementing and Evaluating Search Engines (MIT Press 2010), as well as more than 100 scholarly articles. He serves on the Program Committee of TREC, and also is a coordinator of its Total Recall Track.
The legal marketplace offers many tools, methods, and protocols purporting to employ technology-assisted review (TAR), under names like predictive coding, assisted review, advanced analytics, concept search, and early case assessment. Yet adoption of TAR has been remarkably slow, considering the amount of attention these offerings have received since the publication of the first federal opinion approving TAR use (see Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182 (S.D.N.Y. 2012)). The complex vocabulary and rituals that have come to be associated with TAR, including statistical control sets, stabilization, F1 measure, over-turns, and elusion, have dissuaded many practitioners from embracing TAR.

However, none of these terms, or the processes with which they are associated, are essential to TAR. Indeed, none of them apply to continuous active learning (CAL), the TAR protocol that has achieved the best results reported in the scientific literature to date (see Maura R. Grossman & Gordon V. Cormack, Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery, 2014 Proceedings of the 37th Ann. Int’l ACM SIGIR Conf. on Research & Dev. in Info. Retrieval, 153-62 (2014)).

CAL is a method for finding substantially all relevant information on a particular subject within a vast sea of electronically stored information (ESI). At the outset, CAL resembles a web search engine, presenting first the documents that are most likely to be of interest, followed by those that are somewhat less likely to be of interest. Unlike a typical search engine, however, CAL repeatedly refines its understanding about which of the remaining documents are most likely to be of interest, based on the user’s feedback regarding the documents already presented. CAL continues to present documents, learning from user feedback, until none of the documents presented are of interest.

Counsel interested in using CAL should first become familiar with:

- The promise of TAR.
- The differences between CAL and other TAR protocols.
- How the CAL process works.
- How to determine when a TAR review is complete.
- How to evaluate the success of a TAR review.

**THE PROMISE OF TAR**

In e-discovery, the problem of finding everything that can reasonably be found in response to a request for production has traditionally been addressed through a manual review process, in which every one of a large collection of documents, whether paper or electronic, is reviewed for responsiveness. Often, documents that are unlikely to be responsive are removed using a variety of filters, such as keywords, file types, or date restrictors, before beginning the manual review process. Scientific research indicates that manual review is neither particularly effective nor efficient, and TAR-based methods, where a computer selects only a fraction of the available ESI for review, can be more effective and efficient than traditional manual review methods (see generally Maura R. Grossman & Gordon V. Cormack, Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review, 17 Rich. J.L. & Tech. 11 (2011); Herbert L. Roitblat et al., Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review, 61 J. Am. Soc’y for Info. Sci. & Tech. 70 (2010)).

Most TAR tools use supervised machine learning, where a computer algorithm ranks or classifies an entire collection of documents by analyzing the features of training documents previously classified by the user. The supervised machine-learning algorithms used for TAR should not be confused with unsupervised machine-learning algorithms used for clustering, near-duplicate detection, and latent semantic indexing, which receive no input from the user and do not rank or classify documents.

Supervised machine-learning algorithms that have been shown to be effective for TAR include:

- **Support vector machines.** This algorithm uses geometry to represent each document as a point in space, and deduces a boundary that best separates relevant from not relevant documents.
Logistic regression. This algorithm estimates the probability of a document’s relevance based on the content and other attributes of the document.

Popular, but generally less effective, supervised machine-learning algorithms include:

- Nearest neighbor. This algorithm classifies a new document by finding the most similar training document and assuming that the correct coding for the new document is the same as its nearest neighbor.
- Naïve Bayes (Bayesian classifier). This algorithm estimates the probability of a document’s relevance based on the relative frequency of the words or other features it contains.


CAL VERSUS OTHER TAR PROTOCOLS

A TAR protocol determines how the learning algorithm is used to select documents for review. TAR providers generally employ one of three protocols: CAL, SAL (simple active learning), or SPL (simple passive learning) (see Box, Comparing CAL, SAL, and SPL Protocols for TAR). CAL is much simpler than SAL or SPL, and may be used with any TAR tool, provided that the tool incorporates a supervised machine-learning algorithm that can rank documents by the likelihood that they are relevant.

A number of burdensome steps commonly associated with TAR are absent from CAL, such as:

- Careful crafting of seed sets.
- Determining when to cease training.
- Selecting and reviewing large random control sets, training sets, or validation sets.

(See, for example, Rio Tinto PLC v. Vale S.A., 306 F.R.D. 125, 128 (S.D.N.Y. 2015) (noting that where the TAR methodology uses CAL rather than SAL or SPL, the seed set is much less significant) (see below Completion of a TAR Review.).

Moreover, CAL should achieve superior results, with less review effort than the other protocols. However, it will yield the best possible results only if the TAR tool incorporates a state-of-the-art learning algorithm.

HOW CAL WORKS

To better understand how to conduct a CAL review, Practical Law readers can access a free model CAL system supplied by the authors at cormack.uwaterloo.ca/cal. It contains as a dataset the recently released collection of 290,099 email messages from Jeb Bush’s administration as governor of Florida.

As a running example, consider the task of finding “all communications relating to Apple Inc.” within that dataset. The model CAL system suggests the document shown in Figure 1 as the most-likely relevant document to an initial query using the term “apple.” Other CAL implementations might identify an initial batch of documents containing those shown in Figures 1 and 2, rather than a single document.

Figure 1: Document concerning “Apple Exec. addresses” identified first in response to a CAL search for “apple.”

In this example, Figure 1 is relevant, while Figure 2 is not. Counsel relays this information to the CAL system, which employs a machine-learning algorithm to determine which characteristics of the two documents render Figure 1 relevant and Figure 2 not relevant.

In the model CAL system, this relevance feedback is provided by clicking either the link labeled “Rel” or the link labeled “NotRel” for each document. The learning algorithm might infer that a reference to “Steve Jobs” is evidence of relevance, while a reference to “Mrs. Apple” is evidence of non-relevance. Combining this and other evidence from the two documents and, more generally, from all of the documents reviewed thus far, the CAL system might determine the next most-likely relevant documents are those shown in Figures 3 and 4.

Figure 2: Document from “Jo A Apple” identified in an early batch of responses to a CAL search for “apple.”

Figure 3: “Apple’s Capital Kids Day” document.

Figure 4: Additional document identified by the CAL system after relevance feedback from counsel.
Presumably, these two documents are relevant to the Apple Inc. search, and counsel would provide the CAL system with additional relevance feedback by clicking on the “Rel” links. Generally, the CAL system continues to present similar documents if they are judged to be of interest, learning suggestive terms like “iMovies” as indicia of potential relevance, along with less obvious cues like “Dana Pamella,” “Capital Kids Day,” and “TechNet.”

However, “TechNet” turns out to be a poor choice for indicating relevance. The term eventually prompts the CAL system to suggest the document shown in Figure 5.

By receiving a “NotRel” coding, CAL learns of its poor choice, discounts the term “TechNet,” and avoids further presentation of similar documents.

After suggesting 70 relevant and 38 not relevant documents, the model CAL system will continue to explore the boundaries of relevance until it exhausts all avenues of potential relevance, and suggests increasingly fewer documents of interest. At this point, counsel can have reasonable confidence that they have seen substantially all of the relevant Apple Inc. documents in the dataset.

To increase counsel’s confidence in the quality of the review, they might:

- Review an additional 100, 1,000, or even more documents.
- Experiment with additional search terms, such as “Steve Jobs,” “iBook,” or “Mac,” and examine the most-likely relevant documents containing those terms.
- Invite the requesting party to suggest other keywords for counsel to apply.
- Review a sample of randomly selected documents to see if any other documents of interest are identified.

At some point, however, counsel must decide that they have undertaken reasonable efforts to identify the relevant documents.

**COMPLETION OF A TAR REVIEW**

A problem common to all TAR methods is determining whether or not the review is sufficient.

The CAL protocol has been compared to popping popcorn. After an initial warming period, the kernels begin to pop at a high rate, but the popping eventually slows down and nearly stops. At that point, it is a reasonable assumption that substantially all of the kernels have popped. If the kernels do not begin popping rapidly after a reasonable amount of time following placement of the bag in the microwave, or many unpopped kernels are discovered in the bag after the fact, it is likewise reasonable to assume that something has gone wrong with the popping process and remedial action is needed. However, the process is considered successful if only a small residual of unpopped kernels is left behind.

The same is true for relevant documents that are left behind following a TAR review. The two most critical issues are:

- How many relevant documents have been missed.
- Whether any of those relevant documents are novel or important.
SAL and SPL protocols raise a compound problem that CAL does not: when to stop training the machine-learning algorithm, and how many documents to review once training is complete. To address this problem:

- SAL protocols typically use a randomly selected control set to track the progress of the review.
- SPL protocols typically use ad hoc sampling methods, such as counting the number of instances in which the learning algorithm and the reviewer disagree (overturns), as the basis for determining when to stop. (See Box, Comparing CAL, SAL, and SPL Protocols for TAR.)

Determining whether a learning algorithm is adequately trained has generated considerable controversy. This issue often prompts requesting parties to demand either to participate in or monitor the TAR training process, or to compel the disclosure of
both relevant and not relevant seed set or training documents. Both options tend to be unpalatable to responding parties.

Whether counsel starts out using CAL, SAL, or SPL, counsel should consider training the algorithm using every document that has been subject to human review, after the document has been reviewed. In a SAL or SPL process, a requesting party would be better served by having the responding party feed the results of the final review back into the learning algorithm to identify additional relevant documents, than by litigating the disclosure of the seed set, training documents, and stabilization criteria. This additional training effectively transforms a SAL or SPL protocol into a CAL protocol, thereby providing additional feedback to the learning algorithm.

MEASURES OF SUCCESS

After a TAR review, counsel may wish to determine how many relevant documents were missed in proportion to the total number of relevant documents in the collection. For example, if there were 300 documents concerning Apple Inc. in the Jeb Bush email dataset, and the CAL protocol failed to find 30 of them, the failure rate would be 10%, and the success rate would be 90%. In the information retrieval space, this success rate is known as recall.

Unfortunately, recall is difficult to measure. Indeed, if counsel knew which documents were missed, they would not have been missed in the first place. Where there are few relevant documents in a dataset, it is exceedingly difficult to find them through random sampling, as required for a statistical estimate. It is even more challenging to do so following a thorough review process. (See Maura R. Grossman & Gordon V. Cormack, Comments on “The Implications of Rule 26(g) on the Use of Technology-Assisted Review,” 7 Fed. Cts. L. Rev. 285, 293, 300-12 (2014) (discussing various TAR validation methods).)

Moreover, reasonable minds can differ on the relevance or non-relevance of specific missed documents, particularly on the margins. If, during the validation process, counsel identifies some missed documents, and achieves consensus that they are indeed relevant, counsel can resume the CAL review using the newly identified documents for additional relevance feedback training and thus improve the recall of the review.

Recall may be estimated through the following steps:

■ Count the number of relevant documents found by the CAL system.
■ Estimate the number of relevant documents missed by the CAL system by drawing a random sample of the documents not selected for review, commonly known as the null set or the discards.
■ Calculate an estimate of the total number of relevant documents, which is the number of relevant documents found plus the estimate of the number missed.
■ Estimate recall by dividing the number of relevant documents found by the estimate of the total number of relevant documents.

For example, if the CAL system identified 270 relevant documents and, through sampling, counsel estimates that 30 have been missed, counsel can estimate that there are 300 relevant documents in total and that recall is 270 out of 300, or 90%.

Counsel should beware of using measures that fail to account for both the number of relevant documents found and the number that were missed. Measures counsel should avoid include:

■ Accuracy. This refers to the proportion of all documents that are correctly classified as either relevant or not relevant.
■ Elusion. This refers to the proportion of documents in the null set that are in fact responsive.
■ Overturns. This refers to the number of documents in a sample that are incorrectly classified by the TAR algorithm and are corrected by a manual reviewer.

These metrics are uninformative because it is possible to achieve apparent success while failing to identify meaningful numbers of relevant documents. Using the Apple Inc. example, counsel could achieve accuracy of 99.9%, elusion of 0.1%, and zero overturns — seemingly stellar results — while failing to find a single relevant document.

TRY CAL AT HOME

There is no better way to learn CAL than to use it. Counsel may use the online model CAL system to see how quickly and easily CAL can learn what is of interest to them in the Jeb Bush email dataset. As an alternative to throwing up their hands over seed sets, control sets, F measures, stabilization, and overturns, counsel should consider using their preferred TAR tool in CAL mode on their next matter.

The author, Maura R. Grossman, served as a court-appointed special master in the Rio Tinto case referenced in this article.

The views expressed in this article are those of the authors and should not be attributed to Wachtell, Lipton, Rosen & Katz or its clients.