ASK Catalyst

A User's Guide to TAR

Your Questions Answered About Technology Assisted Review

By John Tredennick, Jeremy Pickens, Robert Ambrogi, Thomas C. Gricks III & Mark Noel

ASK CATALYST:

A User's Guide to TAR

Your Questions Answered About Technology Assisted Review

> *Good information provided by knowledge leaders in the important new field of AI-enhanced document review. Well worth your time to read because many of the basic questions on predictive coding have already been asked and answered.*
>
> **Ralph Losey, Esq.**
> Jackson Lewis P.C.

TABLE OF CONTENTS

Introduction:

Your TAR Questions Answered

Here at Catalyst, we get a lot of good questions about e-discovery technology and, specifically, about technology assisted review (TAR). And we answer every question we get. Whether the question comes from a client, a webinar attendee, or anyone else, we make sure it gets answered.

We have some really smart people who answer the questions. That's not to brag, it's just a statement of fact. We have one of the world's leading information retrieval scientists. We have the lawyer who was lead e-discovery counsel in the first contested case to win approval for the use of technology assisted review. We have a staff brimming with highly experienced technology and litigation-support experts of all kinds.

So early in 2016, we decided to launch a feature on our blog that we called Ask Catalyst. The purpose was to share some of the questions we get and answers we provide. In addition, we invited our readers to submit the questions they wanted answered.

Many of the questions focused on our advanced TAR 2.0 platform Insight Predict and its continuous learning algorithm. Others addressed more-generic TAR topics.

Your questions were so good that we thought it would be useful to compile the questions and answers into a book for handy reference. We hope you find it useful.

Part I

Understanding the Basics of TAR

1

What Is the Difference Between TAR 1.0 & TAR 2.0?

By John Tredennick

Question:

> *Your blog and website often refer to "TAR 1.0" and "TAR 2.0." While I understand the general concept of technology assisted review, I am not clear what you mean by the 1.0 and 2.0 labels. Can you explain the difference?*

Answer:

We developed TAR 2.0 as a shorthand way to describe a new generation of technology assisted review engines that worked differently than TAR 1.0 engines and offered to significantly reduce the cost and time to find relevant documents during a review. While TAR 2.0 is regularly associated with CAL, which stands for continuous active learning, it goes beyond CAL to address a number of shortcomings found in TAR 1.0 products.

Let me start with a description of a typical TAR 1.0 process.

TAR 1.0: One-Time Training

While different products follow different processes, the hallmark of TAR 1.0 is one-time training. In essence, a subject matter expert (SME) codes a control set for relevance and then trains against that control set. When the training is done, the system ranks the remaining documents and orders them by the likelihood of relevance.

Here Are the Typical Steps for a TAR 1.0 Process:

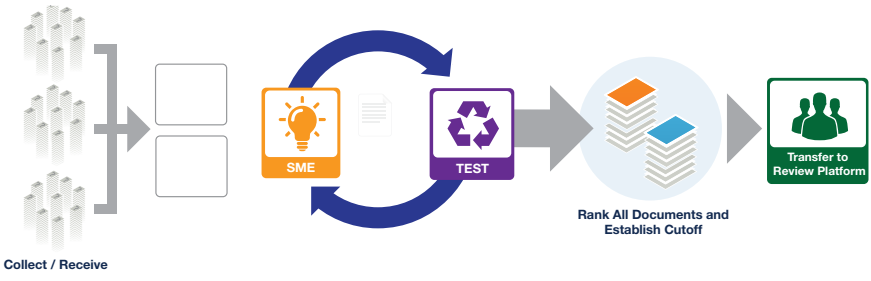
1. An SME, often a senior lawyer, reviews and tags a random sample (500+ documents) to use as a control set for training.

2. The SME then begins a training process using a mix of randomly selected documents, judgmental seeds (documents you find yourself) or documents selected by the computer algorithm. In each instance, the SME reviews documents and tags them relevant or non-relevant.

3. The TAR engine uses these judgments to train a classification/ ranking algorithm to identify other relevant documents. It compares its results against the SME-tagged control set to gauge its accuracy in identifying relevant documents.

4. Depending on the testing results, the SME continues training to improve performance of the algorithm.

5. The training and testing process continues until the classifier is "stable." That means its search algorithm is no longer getting better at identifying relevant documents in the control set. There is no point in further training relative to the control set.



The next step is for the TAR engine to run its ranking algorithm against the entire document population. For testing, the SME might review another random sample of ranked documents to determine how well the algorithm did in pushing relevant documents to the top of the ranking.

Once these procedures are completed, the review team can be directed to look at documents with relevance scores higher than the cutoff point. Documents below the cutoff point can be discarded.

Even though training is initially iterative, it is a finite process. Once the algorithm has learned all it can about the 500+ documents in the control set, that's it. You simply turn it loose to rank the larger population (which can take hours to complete) and then divide the documents into categories to review or not review.

Problems with One-Time Training

When we originally developed Insight Predict, our chief scientist, Dr. Jeremy Pickens, felt the TAR 1.0 process had a number of limitations which he sought to overcome.

1. **One bite at the apple:** The first limitation—and most relevant to a discussion of continuous active learning—is that you get only "one bite at the apple." Once the team gets going on the review set, there is no opportunity to feed back their judgments on review documents and improve the ranking algorithm.

2. **SMEs required:** A second problem is that TAR 1.0 generally requires a senior lawyer or SME for training. Expert training requires the lawyer to review thousands of documents to build a control set, train and then test the results. Not only is this expensive, but it delays the review until you can convince your busy senior attorney to sit still and get through the training.

3. **Rolling uploads:** Another limitation of the TAR 1.0 approach is that it does not handle rolling uploads well, even though they are common in e-discovery. New documents render the control set invalid because they were not part of the random selection process. That typically means going through new training rounds, which is bothersome to say the least.

4. **Low richness:** Low richness collections are a problem for TAR 1.0 because it can be hard to find good training examples based on random sampling. If richness is below 1 percent, you may have to review several thousand documents just to find enough relevant ones to train the system. Indeed, this issue is sufficiently difficult that some TAR 1.0 vendors suggest their products shouldn't be used for low richness collections.

TAR 2.0: Continuous Learning

Working with our technologists and developers, Dr. Pickens created a ranking engine that could rank about 1 million documents in less than five minutes. As a result, it seemed obvious to build a predictive ranking system based on the notion of continuous learning.

Continuous learning means that the algorithm is not limited to one round of training. Rather, as the review progresses, the algorithm continues to learn, taking advantage of the additional judgments made by the reviewers. The reference to "active" means that the system sends documents to the review team based in part on the continuously updated rankings.

CAL's superiority was first documented in a 2014 peer-reviewed study on the effectiveness of the various TAR protocols by Maura R. Grossman and Gordon V. Cormack, who gave CAL its name. They concluded that CAL was far more effective than the one-time training methods used in TAR 1.0 systems. There have been a number of studies since, all concluding that continuous learning is superior to earlier methods.

The process in a CAL review is quite simple.

1. Start with as many relevant seeds as you have or can easily find. These may be documents found through initial searches, through witness interviews or perhaps from earlier reviews. Use these for initial training of the algorithm.

2. Begin the review process. As the review progresses, tagged documents are continuously fed to the algorithm to continue the training. The algorithm continues to rank the documents in relevance order based on the increasing number of training documents.