Replicating Expert Judgment in Document Review

TAR SOLUTIONS FOR A NEW DECADE

PROSEARCH

LOVE WHAT YOU DISCOVER

TAR Solutions for a New Decade

Discovery – for litigation, investigation, regulatory compliance, or any other matters – has evolved dramatically in the past decade or two. It used to be that attorneys and their teams would slog through boxes of paper documents to find relevant materials – think big, dark, dusty warehouses like the ones Mark Ruffalo faced in the film "*Dark Waters*," inspired by the true story of an attorney who took on one of the world's largest corporations in an environmental lawsuit in the 1990s. Following warehouse "collections," teams would then deploy the technology of photocopiers, Sharpies, and Hi-Liters as they organized, annotated, redacted, and sorted materials – think airy, Beverly Hills mansions like the one Alicia Silverstone and Paul Rudd used in "*Clueless,*" inspired by Jane Austen's classic "*Emma*". Today, the same business needs are met with digital data repositories and sophisticated software applications to achieve the same – or better – results.

> Understanding the underlying process and technology is crucial to selecting the right approach for a specific discovery need.

These days modern technology and innovative new tools are used throughout the discovery life cycle, and the change has been so great that the mostly paper processes of yore have been relegated to memory or film (well, digitized film, but that's a different post about technological evolution). Think about it: when was the last time you had a conversation about discovery or document review that didn't include consideration of technology-assisted review (TAR)?

What is TAR?

In its broadest use as a technical term, TAR can refer to virtually any manner of technical assistance – from password cracking to threading to duplicate and near-duplicate detection.

In its narrower use, TAR refers to techniques that involve the use of technology to predict (or to replicate) the decision a human expert would make about the classification or category of a document. In this narrower sense, TAR often comes with a version number – TAR 1.0, TAR 2.0 and, more recently, TAR 3.0.

While some are inclined to advocate for the superiority of a single approach, each version has its merits and place, and understanding the underlying process and technology is crucial to selecting the right approach for a specific discovery need.

This paper discusses some of the variables to consider in choosing the right TAR workflow for a specific matter, as well as the main principles behind different TAR solutions. By doing so, we make the claim that true preparedness lies in (a) understanding the range of core technology within the TAR landscape and (b) further knowing how and where to access the right combination of people, process, and technology to meet any discovery need.

Variables to Consider

The optimal solution to any specific discovery challenge takes into consideration constraints imposed by cost/budget, time, knowledge about the case, and standards for quality, as well as the interaction of these considerations with objective facts about the document set that needs to be searched. Details about the separate, yet interacting, factors to consider include:

TIME: Time considerations include the time it will take to achieve key milestones – starting review, understanding the contents of a document collection, and ultimately production – as well as the time it will take subject matter experts to train a system, when that is required.

COST: Setting aside the hard costs associated with in-house staff as well as attorneys, vendors and document reviewers, considerations include the opportunity cost involved with diverting subject matter experts away from other tasks while they train a predictive model. What's more, some approaches will allow an earlier estimation of the numbers that will help in planning a review as efficiently and cost-effectively as possible, such as the number of documents that will need to be reviewed and the number of responsive documents expected to be found.

KNOWLEDGE ABOUT THE MATTER: The degree to which facts of a matter are known prior to document review impacts the ability to train a model, where that is required. Additionally, prior knowledge may impact how quickly a team needs to have access to "the right documents" to inform both tactical and strategic decisions. The state of knowledge about the case or about information contained in the document collection may impact a team's tolerance for finding surprises in the data relatively late in the review process.

STANDARDS FOR QUALITY: As TAR becomes more prevalent in compliance and discovery arenas, so too does the determination of targets for both precision and recall with respect to satisfaction of discovery obligations. Where minimum thresholds for acceptable quality are known, they can influence the selection of both technology and workflow.

FACTS ABOUT THE DOCUMENT COLLECTION: Some of the important factors to consider about the document collection itself include its completeness – that is, is all of the data that needs to be evaluated available, or is a TAR solution expected to accommodate the rolling ingestion of new data? Additionally, the richness or prevalence of responsive material in a document population can influence the performance of different technologies and workflows and greatly impact time-to-completion.

> TAR can refer to virtually any manner of technical assistance–from password cracking to threading to duplicate and near-duplicate detection.

TAR 1.0

Sometimes called predictive coding, TAR 1.0 or first-generation technology-assisted review solutions leverage examples of both relevant and nonrelevant training documents ("the training set") to train a system to classify documents.

Typically, the training set is coded by a subject matter expert (SME) so the system can replicate an expert's knowledge. How is this done?

- The method used to identify a training set could be based on random sampling (simple passive learning) or uncertainty sampling (simple active learning) or a combination.
- Underlying technologies typically include machine learning algorithms but may also be built on decision trees or complex ontologies of keywords.
- Performance of the system is measured against an answer key ("the control set") that is ideally coded by the same SME that created the training set.

The hallmark of TAR 1.0 solutions is that the training is a finite process that precedes the coding or scoring of all documents. The predictive model and scores associated with it are frozen once training is complete, so changes to either the SME's understanding of relevance or the set of documents that needs to be evaluated require building a new model.

Key Benefits

TAR 1.0 solutions have been demonstrated to outperform traditional linear review, as discussed in *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery* by Gordan Cormack and Maura Grossman[1].

These solutions have the advantage over linear review in that responsive documents are front-loaded during the review process, theoretically providing important information to teams as quickly as possible. Additionally, the control set used in the TAR training process enables an estimation of the number of responsive documents expected to be found, and once the entire population is scored, it becomes clear how many documents will need to be reviewed in total, allowing for teams to gain efficiencies via workflow planning for review.

Criticisms

The main criticisms of TAR 1.0 solutions are based in its "one-time training" feature. One-time training does not allow the system to adjust based on information that is gained during the review. This can be problematic, especially when SMEs in the case are required to code training and control sets prior to synthesizing their expert knowledge of the matter. Because one-time training relies on early coding of a training set, the possibility of bias in the predictive model could introduce concerns about the sufficiency of a production.

ADVANTAGES

TAR 1.0 solutions front-load responsive documents once review starts to provide teams insightful information faster than traditional linear review.

TAR 2.0

In TAR 2.0, the second generation of technology-assisted review solutions, the underlying technique of continuous active learning (CAL) was chosen specifically to improve upon the challenges that one-time training presented for TAR 1.0.

- "Continuous learning" reflects that the predictive model updates throughout the review based on all the coding decisions that humans make, rather than training a model as a discrete step at the beginning of a TAR process.
- "Active" reflects that the system uses the (continuously) updated model to promote the documents with the highest probability of being responsive to the top of the review queue.

TAR 2.0 processes will often supplement these batches of highest-ranked documents with a handful of documents about which the model is most uncertain, enabling it to gather information needed to make better predictions, or documents required to obtain a statistically valid yield estimates, to enable workflow planning efficiencies.

ADVANTAGES

TAR 2.0 solutions require lower upfront training investments than predictive coding and can accommodate continuous data loads.

Key Benefits

TAR 2.0 solutions allow review to begin immediately, without prior training of a model. Additionally, while it may be preferable to have SMEs involved in the early review, this is not a strict requirement, as the model will eventually "smooth over" inconsistent decisions. The low upfront training investment in TAR is considered an advantage over the TAR 1.0 process, especially when this decreases the burden and opportunity cost of having SMEs code documents as part of initial training.

Like TAR 1.0, TAR 2.0 improves upon linear review by presenting a high proportion of responsive documents to reviewers early on. Unlike TAR 1.0, TAR 2.0 can accommodate rolling data loads. And while TAR 2.0 is still subject to challenges presented by low-richness scenarios, it can be prefable over TAR 1.0 insofar as no time is lost waiting to determine whether a predictive model is possible.

Criticisms

There are two main criticisms of TAR 2.0 solutions. The first is that the continuous evolution of the model makes it harder to predict the overall volume of the review. Resource planning in the absence of clear estimates of review size can be inefficient and costly. A "pure" TAR 2.0 approach needs to be supplemented with statistically valid random sampling to help manage this. Even with a yield estimate, the precise timing of how many documents will need to be reviewed ahead of review cutoff is unpredictable.

The second criticism with TAR 2.0 solutions is that the continuous promotion of high-ranked documents introduces the risk of showing reviewers "more of the same" at the expense of promoting diverse responsive content. This introduces the risk of surprises occurring later in the review when unexpected responsive content is lower-ranked.

TAR3.0

TAR 3.0 might better be viewed as the opportunity to combine the advantages of continuous active learning with techniques that minimize the risk of two kinds of surprise: surprises in content and surprises in cost.

1. To minimize surprises in content, TAR 3.0 solutions should be designed to give the system access to a diverse population of documents early in the process. Minimizing surprises only comes with robust knowledge of the document population, and this can be achieved through rigorous sampling and validation.

2. To minimize cost surprises, TAR 3.0 solutions should incorporate methods for determining overall richness to support principled review cutoff and to allow teams to predict the overall volume to enable staffing efficiencies.

The hallmark of TAR 3.0 solutions should be the enrichment of continuous active learning through the application of statistically sound methods of providing early access to the full range of documents, including examples of responsive documents.

A full range of TAR 3.0 solutions could draw on an array of statistically valid sampling methods such as cluster-center sampling (as advocated by Bill Dimm at Clustify[6] and the Clustify blog[5]) or stratification sampling that are augmented with a random sample that is right-sized to enable a yield estimate.

Review commences almost immediately, as with TAR 2.0. But unlike TAR 2.0, the initial review is seeded with documents drawn from targeted samples. A CAL approach continues with the system being continuously updated based on reviewer decisions. Because this approach can incorporate measurable exposure to the breadth of documents, a snapshot of the model can be applied to the unreviewed population to define the outer boundary of a review population to meet case-specific needs.

TAR 3.0 solutions offer the same benefits of TAR 2.0 with respect to minimal upfront investment and the immediate commencement of review. As an improvement over a standard TAR 2.0 solution, the early exposure to a diverse set of documents minimizes the risk of surprises, and the incorporation of yield estimation supports efficiencies in both workflow planning and in identifying a clearly defined review boundary.

Finally, in the case of low-richness scenarios, TAR 3.0 solutions are the superior approach. The ability to front-load examples of responsive documents in a measurable way enables a system to identify responsive documents earlier than in either TAR 1.0 or TAR 2.0 approaches. Pairing this with valid yield estimation is crucial to understanding when review is complete when the prevalence of responsive content is low.

ADVANTAGES

Allowing immediate commencement of review with a minimal upfront investment.

Early exposure to a diverse set of documents to minimize the risk of surprises.

Early yield estimation to support efficiencies in workflow planning and review boundary definition.

Supporting early understanding of conditions required for review to be complete with a built-in method for valid yield estimation.

Superior insight into low-richness scenarios due to an approach that identifies when the prevalence of responsive content is low.

Enriched Active Learning

The newest TAR solutions go beyond delivering a user interface that everyone can work with or one that offers nice-looking graphs and charts for improved data visualization. Rather, they promote more active learning, ensuring the model sees breadth and that samples are reflective of the full range of data to promote measuring decisions.

Enriched active learning is a TAR 3.0 solution that is particularly useful in cases of low richness or when specifics of a case require the identification of a precise cutoff for review. This advanced approach targets the first few thousand documents in the review for seeding with curated samples that ensure a diverse set of documents are presented to a model alongside documents that are likely to be responsive. After these first documents are reviewed, continuous active learning proceeds as in TAR 2.0.

More sophisticated Enriched Active Learning protocols may also incorporate a yield estimate and ongoing reporting, enabling teams to make informed choices about ending review earlier than is possible in most TAR 2.0 protocols.

> More sophisticated Enriched Active Learning protocols may also incorporate a yield estimate and ongoing reporting, enabling teams to make informed choices about ending review earlier than is possible in most TAR 2.0 protocols.

Conclusion

Over the past few decades, the ability to create and store documents digitally has resulted in an explosion of discoverable data, and that has yielded an array of innovative tools for collecting, producing, and reviewing that data quickly and efficiently. Ever-changing technology touches every stage of the discovery life cycle today, and nowhere is that clearer than with document review.

Any conversation about discovery or document review today includes technology-assisted review. As such, it's important to understand the variables and the different solutions available as part of each version of TAR when considering which approach is optimal for each unique matter.

Compared to first- and second-generation TAR solutions, TAR 3.0 can achieve superior results, particularly for teams facing issues with low richness or those under pressure to accelerate review timelines while retaining high levels of accuracy. When considering a TAR 3.0 solution, ask providers about their training methods for active learning.

The TAR Landscape

The modern landscape of technology-assisted review solutions includes three primary categories, typically referred to by their generational introduction: TAR 1.0, TAR 2.0 and TAR 3.0.

A brief overview of the current TAR versions is presented here. The following, more detailed descriptions offer insight on how each is performed along with pros and cons to consider when selecting one approach over the others.

	TAR 1.0	TAR 2.0	TAR 3.0
Does review begin immediately or is it delayed?	delayed	immediate	immediate
Is intense SME review avoided?	No	Yes	Yes
Is a control set avoided?	No	Yes	Yes
Are ongoing reviewer decisions leveraged?	No	Yes	Yes
Can it accommodate rolling data loads?	No	Yes	Yes
Does the approach mitigate surprises?	No	No	Yes
Can the review stop date be estimated early on?	Yes	No	Yes
Does a valid yield estimate fall out of the process?	Yes	Possibly	Yes
How is performance in low-richness scenarios?	Bad	Better	Best

● = Training Activity
● = Reviewing Activity

About The Author

GINA TARANTO

DIRECTOR, APPLIED SCIENCES / ACCELERATED LEARNING SOLUTIONS

Dr. Taranto leads research and innovation of accelerated learning solutions by directing multidisciplinary teams of technologists, subject matter experts, and data scientists to train the technologies that replicate human decisions. She has been developing teams and solutions in eDiscovery for 14 years, with experience in the design and implementation of search and automated document review solutions for clients in the financial services, technology, and pharmaceutical industries. Gina's application expertise includes IBM's Watson Explorer, Equivio Relevance, Relativity, and dtSearch. In addition to building ProSearch's Linguistics, Analytics, and Data Science group, she has led the development of internal programs for happiness, feedback, training, and supporting communities of practice. She is a published author in the fields of linguistics and information retrieval. Previously, Dr. Taranto was a lead linguist and adviser to client engagements at H5, and a research linguist at both A-Life Medical, Inc. and Northrop Grumman Information Systems. She received her B.A. from Kresge College at the University of California, Santa Cruz, with honors, and her M.A. and Ph.D. from the University of California, San Diego.

External Sources Cited

1. *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery* by Gordan Cormack and Maura Grossman
http://plg2.cs.uwaterloo.ca/~gvcormac/calstudy/study/sigir2014-cormackgrossman.pdf

2. Above the Law
https://abovethelaw.com/?sponsored_content=the-evolution-of-predictive-coding

3. E-Discovery Team
https://e-discoveryteam.com/2015/10/11/predictive-coding-3-0/

4. Part of the Solution
https://garywiener.wordpress.com/2015/07/27/on-tar-1-0-tar-2-0-tar-1-5-and-tar-3-0/

5. Clustify Blog
https://blog.cluster-text.com/2016/01/28/tar-3-0-performance/

6. Clustify Blog
https://blog.cluster-text.com/2016/01/28/tar-3-0-performance/

7. Clustify
http://www.cluster-text.com/v/TAR_3_and_training_predictive_coding.php