Technology Assisted Review for the Rest of Us

Caragh Landry
<table>
<thead>
<tr>
<th>TABLE OF CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introduction: Overcoming TAR-Crastination</strong></td>
</tr>
<tr>
<td><strong>Stage 1: Intelligent Batching</strong></td>
</tr>
<tr>
<td>Threading</td>
</tr>
<tr>
<td>Near Duplicates/Near Duping</td>
</tr>
<tr>
<td><strong>Stage 2: Concept Searching</strong></td>
</tr>
<tr>
<td>Keyword Expansion</td>
</tr>
<tr>
<td><strong>Stage 3: Clustering</strong></td>
</tr>
<tr>
<td>Search Results Review</td>
</tr>
<tr>
<td><strong>Stage 4: Find Similar and Categorization</strong></td>
</tr>
<tr>
<td><strong>Stage 5: Predictive Coding</strong></td>
</tr>
<tr>
<td><strong>Stage 6: Prioritized Review</strong></td>
</tr>
<tr>
<td>Conclusion</td>
</tr>
</tbody>
</table>
Introduction: Overcoming TAR-Crastination

“If you want to conquer fear, do not sit at home and think about it. Go out and get busy.” – Dale Carnegie

Someone told me recently that they wanted to use technology assisted review (TAR), but they didn’t know where to start. Immediately, I thought of this Carnegie quote. We still see many clients running linear reviews – batching by custodian and reviewing every document. But with all the technological advancements to the review process in the last several years, this linear approach is too cumbersome, too costly and, frankly, too unnecessary.

My response to those interested in getting started with TAR processes is to begin with the basics and build up from there. I suggest starting with Intelligent Batching and then, once comfortable, add one or more new steps or tools into the process. The following is a six-stage path I would suggest for anyone who wants to ease into the TAR waters rather than cannonballing on the first try – or who might otherwise procrastinate (which I refer to as TAR-crastination).

Stage 1: Intelligent Batching

In a typical linear review process, documents are often organized by custodian and then batched out to reviewers. The reviewers then code each document, one-by-one, and sort them by DOC ID or DOC DATE. If you think about how your own email inbox is set up, this means the reviewers are proceeding without any particular focus, moving from one topic to the next each time they click on NEXT DOC.

By contrast, Intelligent Batching is all about how to better present documents to reviewers, so that the review process is not so haphazard. Every document still gets their eyes on it, but how these documents are presented to reviewers enables them to work more efficiently and code more consistently, too.

Threading

Threading technology keeps email conversations together for the sake of context. This allows a single reviewer to review an entire email chain together, which results in better, more consistent coding of all the emails within each chain. Reviewers will be able to read the email within the context of the whole conversation, enabling them to more easily distinguish the singularly responsive or privileged parts. Without threading, you might have several reviewers seeing only bits and pieces of the conversation, which can lead to coding inconsistencies and even false positives or false negatives. Quality control processes can, of course, catch such issues during a linear review, but with threading, the burden on quality control is greatly reduced.
Near Duplicates/Near Duping

A document that is very similar textually to another document is called a near duplicate. Often these documents can be multiple drafts of the same document, such as PDF copies of an original Microsoft Word document or all the various iterations of an email chain. Several TAR engines on the market today are able to identify near duplicates, which enables Review Administrators to group extremely similar documents together for a more rapid, more consistent review process. By way of example, if you have ten drafts of a single contract, sent to five different custodians, TAR systems can batch all copies, regardless of custodian, into a single batch so that one reviewer can go through and tag them all as responsive, or tag all drafts as not responsive and the final copy as responsive. Reviewing all of these similar documents at once mitigates the risk of inconsistency in producing extremely similar or almost identical documents.

The two batching strategies described, Threading and Near Duping, can enable reviewers to be more focused on topics and thus more consistently code similar documents within those topics. Applying Intelligent Batching to documents is a relatively simple way to help your review team to make better decisions.

Furthermore, if we compare batching by custodian with batching by Threading and/or Near Dupe, in the latter case a reviewer will easily move through their documents approximately 25 percent faster (e.g., going from 75 docs per hour to 94 docs per hour). This could even be a conservative estimate, depending on the amount (and type) of email conversations and the volume of near dupe documents. Intelligent Batching can also increase consistency, which means you can expect to see a 15 to 20 percent decrease in time spent correcting quality control (QC) errors.

The biggest impact in terms of efficiency and speed comes into play with daily or weekly emails and with draft copies of documents. As an example, I have seen reviewers, in less than an hour, go through all 1500 copies of a daily stock report, which are very similar in content. In that same amount of time, I have also seen a reviewer get through more than 100 copies of a document – including various draft versions, sent to different custodians. The main reason these reviewers could work so quickly is because the documents were presented in topically relevant batches. The reviewers were able to focus on specific document types and topics, and view them one right after another. With these advantages, Intelligent Batching is quite simply a better way to organize documents.

Stage 2: Concept Searching

Using key terms to narrow a document population or to find specific, important documents can be a limiting approach, as the key terms only reflect our best guess as to how to find responsive documents, but ultimately it’s still a guess. Over the years, I have seen some really impressive key term lists, chock-full of proximity limiters, double and triple parenthesis and all sorts of wild-card parameters. But, these searches still wind up being either far too broad or far too narrow. Think of key terms searches as akin to fishing for tuna with a trawl net – even if your catch all the tuna, you’ll also spend the majority of your time throwing back the other types of fish you didn’t want.
Key term searching can be a great way to start finding documents, but concept searching is a better method that will help to refine your results and find even more of the relevant documents you need. Concepts are basically just nouns or noun phrases that often reflect themes within documents. Car is a theme, Product Launch is a theme, Mortgage is a theme. However, if you run a keyword search for the term car, you will only find documents that contain the word car. If you run a concept search for the word car, you will find documents about cars, but the word car may or may not actually be in the text. Instead, you might see words like vehicle, truck, Honda or an expression like “my ride.” All of these are often used in place of the word car and in the same context.

By including conceptually similar words, Concept Searching overcomes many of the problems of Boolean keyword searching. The most prevalent issues with Boolean search are:

- Words that have more than one meaning (polysemy)
- Multiple words meaning the same thing (synonymy)
- Misspelled words or typos
- Over/under inclusiveness

Conceptual search technology has the advantage of being able to derive the meanings of words, based on the various concepts in a document, and does not rely solely on identical matches of character strings. It also demands more in the way of investigative skills than dtSearch skills. Overall, the process is pretty straightforward and also agile. You don’t necessarily have to know what’s in your data, so you start with what you do know and build from there. As concept searching increases recall (the total number of returned docs) and decreases precision (the percentage documents that are actually what you are looking for), users can refine their Concept Searching strategies through multiple iterations, drilling closer and closer to what they really want with each refinement. Once you do find some relevant documents, you can expand the result set using options such as More Like This or Find Similar. These tools can even take the full text of a document and use that as a concept search, looking for documents with similar concepts and themes to the one you liked. It is a creative and probing way to find the documents you are looking for in your review.

We also have clients who rely heavily on Concept Searching for their early case assessment (ECA). They begin with some exemplar documents and use the More Like This or Find Similar tool to identify other documents. They will also pull some key phrases and paragraphs from relevant documents and use them in Concept Searching. Since some of the concept searches will be spot-on – and others will return true positives buried among many extraneous false positives – these clients will run multiple iterations of phrases and paragraphs to narrow down the results to the very most relevant ones.

**Keyword Expansion**

Keyword Expansion, which looks for polysemy and synonymy, can be helpful in creating a key term and hot term list. This approach tells you how people in your data set are talking about the same concepts you
input, albeit with different words. A Keyword Expansion example from one of our clients for the word angry yielded additional terms such as mad, miffed, dark, red, crappy, holed, crazy and futile. Sampling these terms and the resulting documents, these were all concepts people were using in their emails to convey that they were indeed angry about something.

The ability to search large collections of unstructured text on a conceptual basis can be more efficient than traditional Boolean searching and, not surprisingly, has become an accepted search methodology in the e-discovery field. We see it most frequently used in ECA, second-level review and with finding exemplar and hot documents. If you’re currently reviewing in a platform that enables Concept Search and Keyword Expansion, I recommend trying them.

**Stage 3: Clustering**

In addition to near-dupe and thread batching, using Clustering for review or search results analysis can be another useful application of TAR. By way of definition, Clustering is the grouping of similar documents based on themes or concepts. Think back to the days when you would receive boxes of paper documents and then would task someone else to sit in a conference room and organize those documents by topic. At the end of the day, you would have a table covered with piles of these printed documents, each pile representing a type or topic, denoted with a sticky-note on top. Sales reports in this pile, brochures and pamphlets in that pile, emails about corruption in another pile, all neatly organized and ready for review.

Clustering mimics this same process, but without human labor and much, much faster too. Based on the text of the documents and the concepts/themes within those documents, Clustering tools create these “piles” of documents for you. After email thread and near dupe batches have been created, batching the remaining population of documents using Clustering will enhance the review team’s ability to make better decisions on the documents, since similar documents will now be grouped together.

**Search Results Review**

Clustering can also be helpful for reviewing search results. Many review platforms on the market today allow you to cluster search results. If you started with search terms, you likely have an idea of what you are looking for in the population, but result sets can be large and often will contain a significant number of non-relevant documents. This is largely due to the limitations of keyword search, as described earlier. By clustering your search results, you create concept folders – or themed piles – of documents for review, each named with the top three to five concepts characterizing the documents in each, i.e., your virtual sticky-notes. This approach will enable you to skip the clusters that seem off-topic and allow you to instead focus on the ones that seem most relevant – just like you would have done with all those piles of documents in the conference room!

Clustering can also be helpful when you have no idea what is in your data set. Even before batching, during ECA or an investigation phase, creating clusters of documents can help identify key themes and groupings of documents for instant insight about the data, all with very little review needed. Simply cluster your
documents to create the grouped “piles” and then, based on what you’re seeing, you can determine which folders most likely contain relevant docs, which ones you can de-prioritize because of a lack of relevance and which, perhaps you can ignore all together (e.g., out of office emails, invitations, spam, etc.).

**Stage 4: Find Similar and Categorization**

Find Similar can have a big impact on finding key and relevant documents within your document population and also in batch tagging (coding) large document sets as not responsive. Find Similar takes an initial identified document or excerpt of text and goes into the document population to find similar documents based on their text content. When you use this feature, you are creating a cluster of like documents from a single exemplar or single exemplar excerpt. Moreover, when you come across an interesting document, positive or negative, you can use Find Similar to return a cluster of similar documents to help you see the impact and depth of the topic or issue.

Additionally, in some platforms, this capability can also help in creating privilege logs. In kCura’s Relativity platform, once you aggregate your documents for the privilege log, there is a useful workflow where you can include similar documents, filter on the Privilege and Responsive fields and use the QC or validate functions to determine why certain documents didn’t make it into the log. This approach can be followed in production QC as well, to find other documents you may want to produce.

Categorization can also play a big role in finding key and relevant documents. While Clustering, as mentioned earlier, is unsupervised machine learning performed without human input, Categorization, on the other hand, allows the user to guide the system and provide exemplars of how to find similar documents. Categorization can be especially helpful when productions are received by opposing counsel or prior to a document request to help organize large document collections for review and production.

**Stage 5: Predictive Coding**

Predictive Coding is a human-driven, computer-expedited approach to document review that allows reviewers to code sample sets of documents and then extrapolate those decisions, based on similarity, to the greater document population. It has been the talk of the town in e-discovery for five or six years now, and every recent LegalTech and ILTA conference has had a track or focus on Predictive Coding. You have also likely heard about the big cases that have utilized Predictive Coding – Kleen, Global Aerospace and Progressive, to name a few – but there are countless other legal teams, operating outside of the media spotlight, who also utilize Predictive Coding workflows as a standard practice in document review and production.

Most review platforms now have standardized workflows that can lead you through the Predictive Coding process. Integreon mostly uses Relativity’s Assisted Review (RAR) and Recommind’s Axcelerate because these two platforms have very structured, transparent Predictive Coding methodologies that are repeatable, explainable and report-heavy.
Statistics can be a scary thing to many people, but they’re an important part of any Predictive Coding workflow. Not understanding what statistics say about the status and progress of a review holds a lot of people back from using these types of workflows. However, most review platforms have solved this issue with automated workflows that can guide you at each step along the way. From the statistics you should care about most to reporting, it is now easier to assess how well you are training the system and when you’ve likely reviewed enough as a sample set. When to stop this initial review is your decision, of course, but the statistics provided – namely precision, recall, F1 and accuracy – will let you know how close you are to capturing all of the responsive documents and how stable your data has become throughout review. The goal of most Predictive Coding exercises is to have high recall and high precision, so that most of the responsive documents in the population are found and for the system to have accurately found them on its own (with few overturns during validation).

From a reviewer perspective, Predictive Coding is rather simple: you review the documents put in front of you. From a manager’s perspective, it doesn’t have to be much harder than that. Most platforms will walk you through the decisions you need to make, including:

- Do you want to create a control (truth) set?
- How will you create your first training set? Using statistical or stratified sampling, or pre-coded seed documents?
- How will you create your QC/validation sets?
- How many rounds of training will you review?
- What is the process for incorporating rolling data deliveries?
- What are the reports saying about your review?
- What do you do when the review statistics plateau?

I am often asked about when to use Predictive Coding and what kind of cases are best suited to this type of workflow. My answer is that most cases are suitable for Predictive Coding workflows. Our clients use it on both small matters and huge matters, financial matters and pharma matters, in small law firms and large law firms. Using technology to assist the review process makes reviews faster, more accurate and more consistent. The cases where we have struggled the most to use Predictive Coding are typically those whose documents don’t have text, such as in intellectual property cases with a lot of technical schemas and drawings, construction cases with a multitude of CAD drawings and securities matters, where there is often more audio than text. Even in these situations, we have been able to use the Predictive Coding tools where appropriate, and route other sets of documents for manual review.

Predictive Coding doesn’t have to be hard and it doesn’t have to be scary. At any point during a review, if you don’t like using Predictive Coding workflows or you don’t think it’s working out, you can always fall back to a traditional review.
**Stage 6: Prioritized Review**

Prioritized Review is the new, hot term in Predictive Coding, replacing it as the buzzword du jour over the last year. In short, to prioritize a review is to address the most relevant documents first. We have all been doing this for years – in our selection of certain custodians as first to be reviewed, in our choice of which boxes (from the hundreds or thousands of them) are most likely to give us what we need, in our decision to look at contracts first, and so on. A Prioritized Review workflow takes those ideas, combines them with Predictive Coding-based machine learning, and extends our reach into the population of documents to find the relevant documents faster and more completely.

A prioritized workflow usually starts with seed documents, which are known relevant documents, and then adds in machine learning, using Predictive Coding technology, to identify other documents with similar relevant content. It’s a continuous process wherein each successive iteration, outputs more and more of the relevant documents to put in front of reviewers to validate. Continually submitting more and more examples of relevant documents to a machine learning process speeds its ability to find most of the relevant content in a document population.

Prioritized reviews further allow you to triage your review team into better, more acute teams of document experts. We often stage a dual-process prioritized review where we have a Relevant team and a Not Relevant team, and then feed the system with seeds of known relevant and known not relevant documents. This allows the management of two separate teams, with two entirely separate goals. The Relevant team reviews for production and focuses on achieving a high recall on relevant documents to get the right documents out the door as quickly as possible. In contrast, the Not Relevant team focuses on finding the junk, unrelated business documents and spam that exist in every review population. We use machine learning to code en masse all those special promotions, personal emails on birthday parties and baseball stats, and other emails that are unrelated to the litigation or investigation at hand. All of this ends up saving us time, and therefore cost, in the review process. While focused reviewers look for the relevant documents, which is very important, the ability to auto-code and batch code the remaining 70%+ of irrelevant documents, which is equally important.

This workflow is also quite conducive to reviews where data collection happens on a rolling basis. By its nature, the workflow is an iterative approach and new data is added into the process in each subsequent round. New custodians or sources often add new concepts (themes) to the document population (for example, documents from a marketing team can be quite different from those created by an engineering team), so having this iterative approach is essential to ‘seeing’ the whole of the data set in your sample documents. By using a prioritized review workflow, review teams can make strategy decisions faster, reduce time and money spent during the review phase and thus more easily meet aggressive deadlines.

**Conclusion**

The six stages described above are the building blocks for a technology assisted review process. Employing a single stage on its own, or combining multiple stages, will allow reviewers to review documents in a
more intelligently organized fashion, with gains in speed and consistency efficiencies throughout. If you don’t have much experience with TAR workflows, start by using just one of these approaches and build up to Predictive Coding or Prioritized Review as your comfort level increases. Be sure to lean on your technology providers and review managers who are experienced with these tools and workflows. They will very likely be happy and excited to guide you. Once again, as Dale Carnegie said, “If you want to conquer fear, do not sit at home and think about it. Go out and get busy.”

About the Author

Caragh Landry is an 18 year veteran of the discovery/e-discovery field and an expert in workflow design and continuous improvement programs for technology assisted review (TAR) and predictive coding (PC) processes, including the design of several review and TAR/PC platforms over the years. In her role on the Executive Management Team at Integreon, she oversees the firm’s Onshore Managed Document Review business and operations in both the US and UK. She frequently consults, speaks and writes on the benefits of TAR and PC, and how these technology-driven processes can be practically applied to create greater efficiency and improved quality in the document review process.

About Integreon

Integreon has more than fifteen years of experience providing hosted review and other e-discovery services for the world’s leading law firms and corporate legal departments. Our services include managed document review provided onshore, offshore, onsite, and including multi-lingual review in more than 50 languages. In over the last five years, we have ensured the successful and timely review of more than 50 million documents in support of a wide range of litigation and compliance matters.
For more information contact:

- Caragh Landry
  Global Head, Managed Document Review
  caragh_landry@integreon.com

www.integreon.com