The Basics of How Technology Assisted Review (TAR) Works – and Why Continuous Active Learning is Best

By Dr. Jeremy Pickens, Catalyst Repository Systems, written for DSi
August 14, 2014

Technology Assisted Review (TAR) algorithms can be daunting. There is no need for the average legal professional to know the complex algorithms and mathematics behind the various TAR engines, but a basic knowledge may clear some things up and increase their comfort level.

The Basics of TAR

At its core, TAR can be thought of as a much larger, more accurate key term search. First, the software makes a list of every word that occurs in all of the documents. Imagine a list of documents on the right and the words on the left, and the software has drawn lines connecting the words to the documents in which they appear. When attorneys mark a document as relevant, responsive or privileged the engine looks at all the terms in that document, uses mathematics to make a determination about the most and least important terms in that document and then ranks unseen documents by the strength of the matching terms.

One may think of this as a sophisticated game of “red rovering.” Remember the childhood game: “Red rover, red rover, send John over?” Once a document is marked as relevant, the software goes over all the words in that document and those words get red rovered, or sent over, to a very long, weighted term list that is then used to search for additional documents that have those words. Then, the software iterates that process so that as new documents are marked responsive, the corresponding new terms are added to the list of search terms. The larger list helps the software predict the documents that are responsive and rank them in order of highest likelihood of responsiveness.

Continuous Active Learning

Some people fear TAR technologies because they do not want to chop off the learning process at an arbitrary point. New TAR tools – like Insight Predict – absolve this fear because they continue to play red rover until there is no reason to continue. These tools use continuous active learning (CAL), which continues to learn with every additional relevance judgment. CAL was shown in a recent Cormack-Grossman study to achieve higher recall than finitely (limited training, stabilization-based) approaches in a recent Cormack-Grossman study. The graphs below from the study visualize this.

Key Term Searches

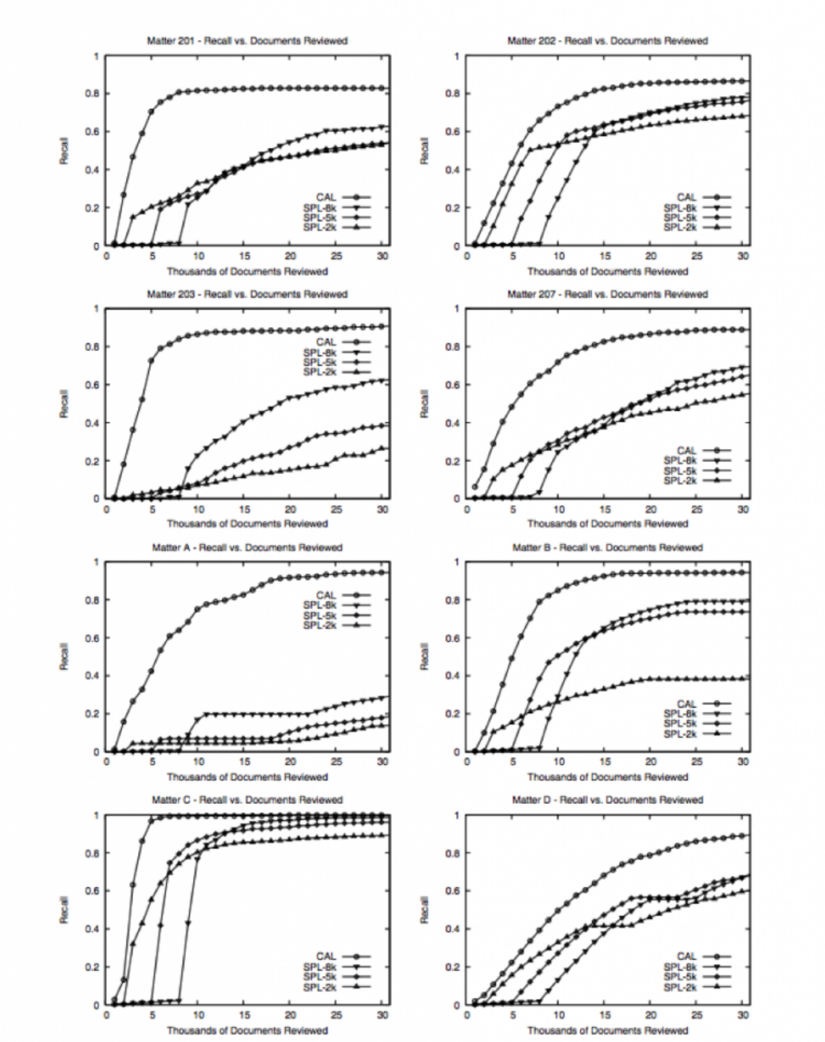

Figure 1: Continuous Active Learning versus Simple Passive Learning using three different training-set sizes of randomly selected documents.

Now, let's compare TAR to key term searches. Key term searches can function in a few ways. Boolean searches bring documents back in random order, while search engines that provide "best matched" will show a ranked list, and most modern systems should have a ranking capability. At the risk of oversimplifying, TAR can be thought of as conducting a large, "best match," ranked result key term search. Once the attorney marks documents as responsive, the software pulls out the key terms and goes back to the remaining list of documents to find additional documents with those responsive key words. TAR uses a huge, weighted term query, but instead of using the dozens of terms a human could pick, it extracts tens of thousands of terms from relevant documents. The computer helps you conduct the search, which is much bigger than any human could handle, by essentially running it in the background and returning to you via the TAR interface the ranking over all the documents in the collection, as determined by this long query. This eliminates the time intensive process of creating the perfect key term query by hand.

And just like Google returns results based on likelihood of relevancy, TAR ranks documents in a prioritized order. Every hit will not need to be examined, as most of the relevant documents will be at the top of the results, and relevant documents that are not initially highly ranked will be found as the process continuously iterates (CAL) with every new relevance judgment made.

Computer algorithms use a tried-and-true process to produce excellent results at a scale that is difficult for humans to reproduce. And when the process guiding the TAR algorithm is one of continuous active learning (CAL), with new knowledge constantly improving the prioritization, there is no need to fear Technology Assisted Review.

Dr. Jeremy Pickens is senior applied research scientist at Catalyst Repository Systems.

To learn more about DSi, visit our blog.
Original article here.

DSicovery®

eDiscovery About People™

About DSi

Serving law firms and corporate legal departments worldwide, DSi is a litigation support services company that provides advanced eDiscovery and digital forensics services. Through five core business processes—DSicollect, DSintake, DSinsight, DSireview, DSisupport—DSi's highly trained staff will help you harness today's most forward technology to gain a competitive advantage. DSi is headquartered in Nashville, Tenn. with offices in Knoxville, Tenn., Cincinnati, Ohio, Charlotte, N.C., Minneapolis, Minn., Philadelphia, Penn., Atlanta, Ga. and Washington D.C. For more information, please visit DSi at www.dsicovery.com or follow us on Twitter at: @DSicovery.