How One Enterprise Dipped A Toe Into Predictive Coding

By: Barry Murphy & Mikki Tomlinson

eDJGroup

Table of Contents

eDJGroup

Introduction

The cost of eDiscovery and document review specifically, has encouraged many corporations and law firms to begin working with Technology Assisted Review ("TAR") solutions. The corporations are interested in leveraging recent analytical improvements to review applications, their law firms are sensitive to higher expectations regarding legal process efficiencies, and legal service providers are striving to provide the brightest reviewers and best-of-breed applications.

One such corporation revealed to eDJ its efforts to reduce legal review costs using a form of TAR known as predictive coding.

Even in conservative estimates, legal document review can be a significant expense. Given the complexity of legal matters, diversity of organizations, and relative newness of digital evidence, there are few generalized statistics on eDiscovery costs. For large organizations, it is not uncommon to have to collect upwards of 100 GB of potentially responsive ESI per matter or investigation. Given typical assumptions about culling rates, amount of documents per GB, amount of documents reviewed per hour, and the hourly cost of review, these companies are looking at legal review costs of over $5 million for each sizeable matter. The legal review costs are staggering, especially for large enterprises that are serial litigants.

Some companies embrace TAR and buy all-you-can-eat solutions so as to use TAR in every matter, regardless of size. Other companies are more cautious, choosing to let case law and more standard TAR practices emerge. One thing is clear: there is no one-size-fits-all way in which to utilize TAR. Serial litigants may chose to bring TAR software in-house and use it on all matters while others will contract with managed services companies on a matter-by-matter basis. Over the coming months and years, anecdotal evidence of TAR usage will build and best practices will emerge.

Background

As explained in eDJ Group's report "Technology Assisted-Review Market Overview," predictive coding is a propagation-based form of TAR. It works by passing along, or propagating, what is known about the matter based on a sample set of documents to the rest of the documents in a corpus. Propagation-based TAR comes in

[1] Assumptions: 100 GB of data with a culling rate of 65% results in 1,950,000 items for first-pass review; first-pass review with analytics at 250 documents per hour with hourly rate of $75/hour; 30% of documents make it to second pass review where rates are 50 documents per hour at cost of $400/hour.

different flavors, but all involve an element of machine learning. In some scenarios, a review team will have access to a "seed set" of documents that the team codes and then feeds into the system. The system then uses the seed set to mimic the action of the review team as it codes the remainder of the corpus. In other scenarios, there is no seed set; rather, the systems give reviewers random documents for coding and then create a model for relevance and non-relevance. It is important to note that propagation-based TAR goes beyond simple mimicry – it is about creating a linguistic mathematical model for what relevance looks like.

While there are different ways to execute predictive coding, the goal is always the same: to combine technology with people to speed, improve, and potentially automate some elements of the legal review process in such a way as to reduce costs and improve quality. It is important to note that predictive coding is not simply about technology; rather, it is about the right mix of technology, people, and workflow. That mix may be sourced from a variety of constituents: within the organization; from external law firms; or from external service providers.

The Case Study

The company, a Fortune 100 company with more than $50 billion in revenue, is no stranger to eDiscovery. On average, it faces 50-100 requests for ESI per year due to litigation, regulatory requests, or internal investigations. The size of each matter can range greatly, from just a few gigabytes to hundreds of gigabytes. The company is always interested in better managing eDiscovery, but its Legal department – like most – is cautious about adopting new practices. Thus, predictive coding was of interest to the company, but the situation would have to be just right. The ideal scenario would need to allow for testing and measurement of predictive coding while also not injecting a high amount of risk into the matter. After all, this company had no experience with predictive coding and did not want to have the defensibility of the review process called into question.

The Legal team at the company continuously monitored potential matters for the ideal case on which to test predictive coding. The situation arrived in the form of an internal matter where the company wanted to understand the facts of the matter to quickly identify and interview witnesses. Because the matter was purely internal – not a civil or criminal case, the company would not have to defend the

Language Plays a Role

Because this particular matter crossed borders, 80% of the document corpus contained non-English language content. A traditional linear review would have required a large team of foreign language reviewers and required a lengthy review schedule. The review costs could have been astronomical. By using Predictive Coding instead of traditional linear review, the Legal team knew what they needed to know within two weeks. Foreign language reviewers were still required for looking at some of the documents, but the team of five reviewers was able to review and translate the hot documents on the fly to get the facts to the legal team faster.

predictive coding protocol in court. That eliminated some of risk. And, because time was of the essence and there was a desire to interview witnesses on an expedited schedule, this matter was a perfect opportunity to test whether or not predictive coding could deliver the relevant facts of the case very quickly. Additionally, the corpus of the collection included documents in multiple languages, which enabled the Legal Team to assess the foreign language capabilities of the TAR platform.

The company has a collection and first-pass processing engine in-house and relies on service providers for eDiscovery activities further downstream in the lifecycle. For this matter, which resulted in a collection of 300 GB of data, the company decided to leverage the predictive coding capabilities of a national service provider that could bring both technology and human expertise to the process so that the corporate reviewers could focus on what they needed to do, which is review the documents.

The company was unsure of what exactly it was looking for, so there was no way to do much culling before sending the data out for processing and review. The service provider loaded the 300 GB of data into the predictive coding solution. A reviewer from the corporate legal team reviewed 40 documents, highlighting the strings of text that made the document relevant to this specific matter. Based on those 40 documents, the rest of the documents in the collection were assigned a weight, with a higher weight indicating a higher probability of relevance.

Now, the rest of the review team was set to begin reviewing documents. The documents were presented to reviewers in order of weight and the team kept reviewing documents until they began finding no relevant documents any longer. As the eDiscovery team leader said, "after weights dipped below 95%, the relevance fell off a cliff." To be doubly sure that the predictive coding process was working, the company took the initial seed set of 40 documents and augmented it with the work product of the review to date. The reviewers then looked at the remaining documents in order of weighting again; after the top 5%, there were virtually no relevant documents. This gave the team confidence that the predictive coding process was finding all the relevant documents and not missing any.

This process allowed the company to get to the most relevant documents early in the case; this helped the Legal team to interview witnesses in the early stages of the matter with a better understanding of the issues. Outside counsel, however, was not fully comfortable with the predictive coding process and proposed some targeted searches of the data based on keywords. The company's Legal team wanted to convince outside counsel that the predictive coding process was, in fact, sufficient and that additional searches were unnecessary. To do this, the company and outside counsel developed targeted keyword searches to run against he unreviewed population of documents. The search resulted in approximately 2,500 documents responsive to the narrowly-tailored keywords that had not been reviewed during the predictive process. When all of those 2,500 documents were deemed non-relevant, outside counsel was comfortable that the predictive coding process was accurate and complete.

The Results: Cost Savings, Improved Quality of Review, and Getting To The Facts Faster

Had the company needed to conduct traditional linear review, it would have been required to review 70,000 documents. With predictive coding, only 20,000 documents were reviewed. Given standard assumptions (a reviewer can review 60 documents per hour and the blended average cost of review is $150 per hour), that adds up to a savings of $125,000 for one small internal investigation. Extrapolate that out to a larger cases and consider that this is just one of hundreds of matters that occur over the course of a year and there are significant savings to be had.

In addition to cost savings, the Legal team at this company pointed to improved review quality as an important benefit. To this team, the keyword method is imperfect. In the context of an investigation like this one, where reviewers are not exactly sure of the facts that they are looking for, there is a tendency to go on a fishing expedition and resist narrowing search terms and being over-inclusive on relevance.

The legal team believes that predictive coding is a big step up because it gets to the facts more quickly. When a reviewer is not exactly sure what he/she is looking for, the ability to have documents prioritized based on the review results from a small seed set can speed up the process of figuring out exactly what is going on in a matter. In this case, the company was able to quickly determine the facts of the case, interview witnesses, and head off a situation before it could become a major litigation problem.

This company is not yet ready to use predictive coding on every single matter; it will continue to evaluate it on a case-by-case basis. That said, TAR is now on the menu of items the company will consider in any case. And, the Legal team sees the opportunity to leverage TAR more strategically as time goes on and comfort level rises. The company is beginning to create case repositories so that that work product can be reused; the Legal team envisions using predictive coding to pull down documents from the master repositories by issues as opposed to keywords. When there are new collections, the master repositories can serve as seed sets to help determine privilege and relevance before doing further work on the collection.

Lessons Learned

There is not a single, standard way to undertake advanced forms of TAR such as predictive coding. Corporate attitudes about TAR vary as greatly as the personalities of the General Counsel. For every company that embraces TAR whole-heartedly, buying all-you-can eat software licenses and using TAR on every matter, there is a company that feels TAR is not yet ready for prime time. For every company that believes on-premise TAR software is the right way to take control of legal review costs, there is another that is on the managed services bandwagon.

The lack of TAR best practices and standards can be frustrating, but eDiscovery professionals should take solace in knowing that each emerging case study continues the evolution of TAR. Even in this one small example, there are valuable lessons to take away. First, small project experimentation can lay the groundwork for larger TAR projects. This case study is proof that even very large enterprises with high litigation profiles can be cautious with new review methodologies. While the potential cost savings associated with TAR are hard to ignore, companies also need to have a solid comfort level with the defensibility of the process. For this company, learning how predictive coding works on a small, low-risk matter creates the necessary comfort level to begin using it on larger matters going forward.

This case study also teaches that legal insight is as big a TAR benefit as reduction of review costs. Much of the hype around TAR focuses on the potential to drastically reduce legal review costs. For this company, getting to the facts of the matter quickly, especially since reviewers were not sure exactly what they were looking for, was invaluable.

About the eDJ Group:

eDJ Group offers unbiased information and pragmatic advice, based on years of experience and proven industry best practices. Whether researching a technology or service solution, conducting an eDiscovery Bootcamp or finding the right expertise to answer your specific questions, eDJ Group is the source for all eDiscovery professionals.

We are committed to helping eDiscovery professionals get the information necessary to excel in their professions, rather than offering legal advice or counsel. We operate with the utmost integrity and commitment to our clients on these guiding principles:

- Independence – All research, reports, advice and services are agnostic and conducted independently without influence by sponsors.
- Highest Ethical standards – All content is honest perspective based on real experience and interactions with thousands of practitioners; detailing both successes and failures without favoritism.
- Pragmatic, Experienced Expertise – All services are conducted by industry experts with decades of experience in eDiscovery and strictly vetted by the eDJ Group founders.

For further information about the eDJ Group and their research, please contact Barry Murphy (barry@edjgroupinc.com) or Jason Velasco (jason@edjgroupinc.com).