THE REVIEW OF
SECURITIES & COMMODITIES REGULATION
AN ANALYSIS OF CURRENT LAWS AND REGULATIONS AFFECTING THE SECURITIES AND FUTURES INDUSTRIES

Vol. 46 No. 3 February 6, 2013

PREDICTIVE CODING: SILVER BULLET OR PANDORA'S BOX?

The high costs of e-discovery have led to the development of computerized review technology by which the user may search for documents by concept rather than by words used. The technique begins with an attorney reviewing a "seed set" of documents and then using "analytics" technology to extrapolate to the larger set. The author outlines various uses for such automated review, the circumstances in which it will be cost-effective, and the limitations of its benefits.

By Jill Crawley Griset *

If you handle litigation or government investigations that implicate electronic discovery, you no doubt have heard of "predictive coding" or "automated review."[1] When these terms are used by practitioners and the courts, they typically refer to a process that involves reviewing a sample data set (sometimes called a "seed set"), having a knowledgeable attorney tag those documents (usually for relevance), and then using "analytics" technology to extrapolate the attorney's decisions to the larger set of documents that have not been actively reviewed.

Think of it as Pandora (the Internet radio application) applied to e-discovery – just as Pandora decides what music you may like based on your previous song selections – the analytics technology identifies documents likely to be relevant or not relevant based on the attorney's tagging decisions in the "seed" set. The allure of this technology is that it appears to dramatically reduce the high cost of e-discovery. Now, because a computer can do the review based on the lawyers' work on the seed set, the days of hundreds of contract attorneys reviewing documents for weeks and months are no more. Because it appears to offer a viable solution to skyrocketing e-discovery costs, it has received the attention of lawyers, large companies, and, more recently, the courts.[2]

[1] As used in this article, the terms "predictive coding," "automated coding," and "automated review" are intended to refer to the process of using analytics to apply decisions made on a sample set to a larger group of documents. None of these terms are intended to refer to or to endorse any specific vendor's product.

[2] *See, e.g., Da Silva Moore v. Publicis Groupe & MSL Group*, 2012 WL 607412 (S.D.N.Y. February 24, 2012); *Kleen Products, LLC, et al. v. Packaging Corporation of America, et al.*, 2012 U.S. Dist. LEXIS 139632 (N.D. Ill.

* *Jill Griset is a partner at McGuireWoods, LLP in Charlotte, North Carolina. She co-directs the firm's Discovery Counsel Services Group, serves as National Discovery Counsel for several large financial services organizations, and manages the firm's Charlotte e-discovery facility. She can be contacted at jgriset@mcguirewoods.com.*

IN THIS ISSUE

RSCR Publications LLC Published 22 times a year by RSCR Publications LLC. Executive and Editorial Offices, 2628 Broadway, Suite 29A, New York, NY 10025-5055. Subscription rates: $1,197 per year in U.S., Canada, and Mexico; $1,262 elsewhere (air mail delivered). A 15% discount is available for qualified academic libraries and full-time teachers. For subscription information and customer service call (866) 425-1171 or visit our Web site at www.rscrpubs.com. General Editor: Michael O. Finkelstein; tel. 212-876-1715; e-mail mofinkelstein@hotmail.com. Associate Editor: Sarah Strauss Himmelfarb; tel. 301-294-6233; e-mail shimmelfarb@comcast.net. To submit a manuscript for publication contact Ms. Himmelfarb. ISSN: 0884-2426.

While at a high level predictive coding seems like a good idea, there are substantial and often unappreciated hidden costs. The concept that Pandora used so effectively for Internet radio may become a "Pandora's Box" if applied to e-discovery indiscriminately by courts and lawyers who have not used the technology. To prevent this, litigants and courts need to understand that predictive coding is just one tool, but not the only tool, for implementing full and fair disclosure of relevant materials in a cost effective way. This article explores the situations when predictive coding may be best utilized and the factors attorneys should consider when evaluating when and how to use it.

EFFECTIVE USES FOR PREDICTIVE CODING AND ANALYTICS TECHNOLOGY

The technology behind predictive coding is often referred to as "analytics." There are numerous features of the technology and it is offered in different forms depending on the vendor used. The technology usually includes the ability to group together documents with similar "concepts" and allows the user to search for documents by "concept" rather than by a particular word used. To do this, the technology runs an algorithm across the documents that identifies terms that usually appear together or "concepts" that appear in relevant documents. Unlike search terms, which are limited to the particular word, this form of analytics can locate documents in which a person may be discussing a trip to New York even where they do not mention the word "New York." For example, the tool may notice other related words in the data set that are often associated with New York, such as "Times Square" and "vacation" and "Statue of Liberty," etc. This allows the user to search for the concept, "New York," and retrieve documents that relate to New York but may not mention the name. Predictive or automated coding brings this technology one step further, by taking groupings of documents tagged as relevant by lawyers familiar with the case and figuring out whether other documents in the larger data set have similar characteristics or concepts to the tagged documents.

Different aspects of the technology can be used very effectively even if true "automated coding" is not used in the strictest sense. For example, a party may

- use analytics on a sampling of data to test and validate search terms;
- use clustering, concept searching, or similar analytics technology on the culled data set to help find sets of documents that are clearly nonresponsive after search terms have been applied to avoid review of those documents;
- use analytics in the review to feed logical groupings of documents to the reviewers, which can accelerate the review significantly;
- use analytics to check work of reviewers for accuracy; and
- use analytics to find documents similar to (or with the same concepts as) specific documents previously identified as important to the case.

True automated or predictive coding (*i.e.*, automatically tagging a group of documents based on decisions made by attorneys on an earlier sample set) may be most cost-effective in the following circumstances:

- internal investigations, in which there is generally a need to find an answer quickly and no opposing party with whom to negotiate how and to what extent predictive coding may be used;
- litigation in which both parties have potentially high volumes of electronically stored information ("ESI") to collect and produce; and

footnote continued from previous page…

Sept. 28, 2012); *Global Aerospace v. Landow Aviation, L.P.*, 2012 Va. Cir. LEXIS 50 (Va. Cir. April 23, 2012); *see also* Transcript of Hearing on Motion for Summary Judgment at 66, *EORHB, Inc et al. v. HOA Holdings, LLC, et al.*, Case No. 7409-VCL (Del. Ch. October 15, 2012)(ordering, *sua sponte*, that the parties use predictive coding or to show cause why it was not appropriate in the case if they wished not to use it).

- litigation in which one side has a high volume of documents, but uses the technology to prioritize the review and production of documents to identify the most important documents quickly, but does not otherwise exclude sets of documents from review.

THE HIDDEN COSTS OF AUTOMATED CODING

Why not use predictive coding in every case? The case of *Da Silva Moore* highlights the difficulties that may arise, particularly in contested litigation where one side has a large volume of ESI.[3] In *Da Silva*, the plaintiff class alleged gender and pregnancy discrimination claims against Publicis Groupe, an advertising conglomerate. While the parties agreed that predictive coding may be used in the case, they disagreed on the details of *how* it should be used. The parties, therefore, had a number of conferences to discuss methods of culling and reviewing the large amount of data. The defendants originally proposed reviewing and producing only the "top 40,000" documents – *i.e.*, the 40,000 documents that the tool identified as most responsive. The court rejected that proposal as a "pig in a poke" and held that "where [the] line will be drawn [as to review and production] is going to depend upon what the statistics show for the results."[4] The parties then spent several months arguing about how the predictive coding process might proceed. Finally, after a hearing on February 24, 2012, the court entered an order approving a method to conduct the predictive coding.[5] The plaintiffs then filed a Motion for Recusal and Disqualification, arguing, in part, that the presiding judge, Judge Peck, should be recused due to his public comments in support of and alleged bias toward predictive coding.[6] That motion was denied.[7]

At the end of the day, the parties spent at least *six months* fighting about the protocol to use to cull/review the documents. The defendant ultimately agreed to search *30* custodians' data as a "first phase" over at least a period of three years. The parties also agreed to search other sources, such as shared folders.[8] The defendant stated that at the time of the February 2012 hearing, it had paid $350,000 in e-discovery costs and estimated an additional $200,000 to pay for additional e-discovery related activities using predictive coding.[9]

The experience of the parties in *Da Silva* shows that while at a high level predictive coding may appear to be more cost effective, the cost of negotiating the parameters of predictive coding is high and may eclipse the cost of a traditional review. Furthermore, the transparency required in that process (at least as interpreted by Judge Peck) makes it very difficult to reach agreement without court intervention. The court in that case, for example, required the producing party to give its entire seed set to the other side, including nonresponsive documents. The producing party is likely not going to be comfortable with this. Further, other issues are likely to arise as the parties haggle over numerous details in setting up the analytics index, the tagging structure, and other details.

By contrast, because search term filtering is more accepted and there are fewer factors to negotiate, the negotiation costs are lower and it is less likely that the parties will need court intervention. With search term culling, the parties do not generally discuss how the producing party is choosing documents that produce search terms or how it is tagging documents – *i.e.*, the producing party does not provide its review guidelines to the opposing party and usually the technical details of the search are not discussed. As a result, there are fewer issues to negotiate and the parameters of the search can often be negotiated quickly and cost-effectively.

Lawyers considering predictive coding should consider the following questions when determining whether to use the technology:

- Does one party have a disproportionate amount of documents? If so, and the opposing party lacks experience with the technology, it is likely that the negotiating costs will be higher if predictive coding is proposed, as search terms are more accepted and easier to negotiate. The plaintiff's attorney is likely to require a high level of transparency in the process with predictive coding that is likely to slow the process and increase costs, and because the plaintiff does not have a high volume of documents, they will not appreciate the costs of the technology. Search terms, however, are likely to be quicker and less costly to negotiate.

- Is the case one where both parties have a lot of documents to produce? The cost of negotiating

[3] *Da Silva Moor*, *supra*, note 2.

[4] *Id.* at *3.

[5] *Id.*

[6] *See Da Silva Moore v. Publicis Groupe & MSL Group*, 2012 WL 2218729 (S.D.N.Y. June 15, 2012).

[7] *Id.*

[8] *Da Silva Moore*, *supra* note 2 at *4-5, 13.

[9] *Da Silva Moore*, *supra* note 6 at *21.

predictive coding parameters will likely be much less because both sides will have an incentive to reduce costs and avoid a large document review.

- What is the size of the data set? Predictive coding is typically much more expensive per gig than traditional search term culling. Thus, the attorneys should consider that cost and whether it may make sense to initially cull the data using search terms to reduce the size of the data set that will need to be indexed for analytics.

- Will the use of predictive coding cause the original data set to expand? Believing that predictive coding is a panacea, courts may permit (or requesting parties demand) wider ranging collection, and exercise less discipline targeting the collection to a narrower scope of sources (both data sources and people). Litigants should be encouraged to do the hard work of identifying truly relevant custodians up front and not be relieved of that burden.

- In a government investigation, is the government asking the producing party to turn over all of its unfiltered data to run through an automated coding tool? Confidentiality and privilege issues will likely arise and the producing party should think very carefully about producing documents without reviewing them. In addition, the predictive coding tools are less successful at identifying nuanced privilege calls than they may be at clearly defined, broader relevance issues.

- Will there still be costs to review the data that will be produced and what will those costs be? It is likely that even if the parties use predictive coding, they will want to review the documents that are identified as relevant before production. Thus, in the typical, carefully managed case, lawyers and clients will be reluctant to view predictive coding as a replacement for active review. Review costs are rarely eliminated.

- What are the estimated costs for "testing" the technology? The producing party will likely still have to spend time quality checking or "QC-ing" the documents that were identified as relevant (or not), just as QC is used in a traditional contract attorney review. These costs must be included in the discovery budget.

- Can analytics technology be used in other ways to reduce the costs even if "automated coding" in the strictest sense is not used?

CONTROLLING THE COST OF DISCOVERY – A MULTIFACETED APPROACH

Before the parties and courts decide that predictive coding is appropriate in a given case, they should take a hard look at the costs and consider whether there are other ways to locate a targeted, relevant set. Predictive coding should not be a substitute for requiring the parties to cooperate in determining a *reasonable* ESI collection involving a small set of key custodians, data sources, and a discrete time period. Often the greatest cost savings will be realized by focusing, at the beginning of the case, on the most likely relevant sources of information (people and data). To do so, the parties should first carefully examine the discovery requests and determine whether it is necessary to collect and review e-mail data to find the core documents responsive to each of the requests. E-mail often consists of day-to-day conversations that will never be seen by a witness or a jury. The courts should be pressing the government and private parties to be reasonable when seeking e-mail and to think carefully about (1) the core issues in the case that might be best addressed by e-mail; (2) the small group of key people likely to have those important e-mails; and (3) the discrete time frame that is likely to contain the most relevant e-mails.

Some courts are, in fact, limiting a party's ability to seek large volumes of e-mails in discovery, at least in certain types of cases. The Eastern District of Texas, for example, has issued a model order in patent cases that seeks to limit the parties' ability to seek large volumes of e-mail data.[10] The order provides that "[e]-mail production requests shall identify the custodian, search terms, and proper time frame," and that each requesting party "shall limit its e-mail production requests to a total of eight custodians per producing party for all such requests."[11] While some may argue this is unique to patent cases, it should be applicable in many other cases as well.

[10] *See* Model Order Regarding E-discovery in Patent Cases, United States District Court for the Eastern District of Texas, http://www.txed.uscourts.gov/page1.shtml?location=rules (Appendix P); *see also* the Model Order proposed for patent cases by Chief Judge Randall R. Rader, United States Court of Appeals for the Federal Circuit *at* http://memberconnections.com/olc/filelib/LVFC/cpages/9008/Library/The%20State%20of%20Patent%20Litigation%20w%20Ediscovery%20Model%20Order.pdf.

[11] Model Order Regarding E-discovery in Patent Cases at ¶ 8, *supra* note 10.

In accordance with this approach, and with the goal of focusing discovery efforts on the most likely sources of important information, the parties should focus on creating a small list of key custodians who are most likely to have important information. Lists of people who are likely to be corresponding with each other should be whittled down to focus on the one or two who have the most important or comprehensive set of documents for each core issue in the case. The lawyers can then create a budget on a per-custodian basis by collecting and examining the volume of e-mail collected from one or two custodians for the requested (or agreed upon) time period, and estimating the total volumes based on those samples.[12] The producing party's lawyers should also talk to the key custodians and other witnesses involved in the matter to confirm which people are key to the issues relevant to e-mail data.

Finally, the producing party should map out the costs with each option and determine whether, at the end of the day, predictive coding will in fact save money. In most large cases, some use of analytics technology in the process will result in cost savings. The trick is figuring out where and when to use it to maximize the cost savings and when its use will lead to protracted, expensive battles. Ultimately, the use of analytics and predictive coding may become as generally accepted as search terms are now, but we are not there yet. To really use the technology effectively and prevent it from becoming a "Pandora's Box," it is important that the courts and the parties understand all of the costs and the limitation of the technology itself, and employ a multi-faceted approach that combines analytics technology with a reasonable, targeted ESI collection.■

[12] Although other electronic data may be relevant, the ESI that typically generates the most cost and highest volume is e-mail, and, because it is typically unorganized, it creates the most challenges to searching and producing.