New York Law Journal

A NEW YORK LAW JOURNAL SPECIAL SECTION

An ALM Publication

E-Discovery

WWW.NYLJ.COM

MONDAY, MARCH 19, 2012

Predictive Coding Is a **New Tool** In the E-Discovery Toolbox

BY PHILIP COHEN AND LAUREN HARRISON

TECHNOLOGY has led to an explosion in the amount of electronically stored information (ESI) maintained by corporations and individuals. Litigators and legal departments face the challenge of keeping the costs associated with discovery under control while avoiding potentially crippling sanctions for mishandling ESI. Current strategies to contain e-discovery costs include limiting the number of custodians and data sources processed; using technological tools such as keyword searches and concept searches to cull down the universe of potentially responsive data to be more manageable; and employing contract attorneys to review and code each document at a fraction of outside counsel's standard rates. Recent developments in technology-assisted review, however, present an attractive option to comprehensive manual review, offering the promise of a more efficient, less-expensive process, and more accurate results.

Attorneys being pitched predictive coding[1] tools by litigation vendors (and the pitches are flying fast and furious—"It's easy!" "Cut your e-discovery costs by 90 percent!") are hesitant to incorporate predictive coding technology into their e-discovery protocols. The most commonly cited reason for attorneys' reluctance to use predictive coding technology is the uncertainty of judicial acceptance,[2] as attorneys are loathe to recommend their clients invest in ESI predictive coding protocols and the related processing and consulting fees without clear judicial authority that such a review is reasonable and defensible.

iSTOCK

Well, the wait for judicial guidance is over. On Feb. 24, 2012, Magistrate Judge Andrew J. Peck of the U.S. District Court for the Southern District of New York, a thought-leader in the ESI field, issued, for "the benefit of the greater bar,"[3] the first opinion that approves of the use of predictive coding.[4] Judge Peck concluded that "computer-assisted review is an available tool and should be seriously considered for use in large-data-volume cases where it may save the producing party (or both parties) significant amounts of legal fees in document review."[5] With this fresh stamp of judicial approval, predictive coding technology looks to be the future of e-discovery. It's time for the bench and bar to embrace the future.[6]

Introduction to Predictive Coding

Predictive coding is a "machine learning" process that requires the involvement of both humans and computers to identify potentially responsive documents. The first step in the process is for the producing party to identify a random sample of documents, which will be reviewed for relevance, to develop the baseline for calculating recall and precision in the review. ("Recall" and "precision" are widely used measurements of document retrieval effectiveness. "Recall measures how well a system retrieves all the relevant documents; and precision, how well the system retrieves *only* the relevant documents."[7]) The parties will then establish additional samples, usually targeted samples identified with keyword searches or samples identified by the client and counsel as target populations of documents known to be responsive. These samples taken together will be used as a "seed set" to provide the initial training of the system to identify potentially responsive documents.

The "training" of the system is best conducted by an identified, accountable, senior case attorney with in-depth knowledge of the case's facts, legal issues, discovery pleadings and case management protocols. A predictive coding application will generally show the reviewer how the system is coding the documents, and the system is considered trained once the human reviewer and computer are sufficiently in agreement such that the rate of responsive to non-responsive documents being identified by the system is appropriate. Predictive coding tools employ algorithms to review vast sets of records after the system has

PHILIP COHEN *is co-chair of Greenberg Traurig's national e-discovery and e-retention practice in New York.* LAUREN HARRISON *is an associate at the firm.*

been trained, on a document-by-document basis, by only a few thousand documents.

Once the "seed set" of documents has been fully reviewed, the system will then retrieve numerous samples of documents identified as potentially responsive. The attorneys will then conduct a number of iterative reviews aimed at further refining the initial responsive population. This iterative process should continue until the attorneys are confident that the system has largely recalled the responsive population. For quality control, attorneys can review the unselected or lowly ranked population, usually by sampling a portion of those documents.

While some tools code the documents as responsive/nonresponsive, other tools rank the potential relevance of the document on a scale of 100 to 1 coding the entire universe of potentially responsive records. The predictive coding tool can then sort the most-likely responsive documents up front, allowing counsel to make a judgment that after a given degree of certainty, it is not efficient for attorneys to review the remaining records (e.g., attorneys should not have to review 100 documents to find one relevant document, but rather should only spend their time and resources reviewing "target rich sets" of documents rated most likely to be responsive).

Advantages of Predictive Coding

Research has shown predictive coding technology can be utilized as a cost-saving, time-efficient, and accurate e-discovery tool. The results of a recent survey taken of 11 e-discovery vendors by The Electronic Discovery Institute found "on average, predictive coding saved 45 percent of the costs of normal review."[8] This number was calculated even "beyond the savings obtained" by current cost-saving measures such as "duplicate consolidation and e-mail threading."[9] Further, seven respondents reported savings of 70 percent or more "in individual cases."[10]

Cost savings can be found with the speed in which a predictive coding tool can code a large population of records. A predictive coding tool can be trained to cull and code a one million document population in a week. For attorneys facing short timelines and huge volumes of data (and it seems like all of us are), it is easy to see the appeal of these tools.

Notably, studies have found that predictive coding technology, properly used, can be more accurate than traditional manual review. In a recent article titled "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review," the authors set out to refute the hypothesis that "manual review is the best approach by showing that technology-assisted review can yield results that are more nearly complete and more accurate than exhaustive manual review[,]" and their analysis found just that.[11] The authors compared the results of five technology-assisted reviews with the results of five manual reviews conducted as part of the TREC Legal Track Interactive Task.[12] The results of the comparison demonstrated that "the average efficiency and effectiveness of the five technology-assisted reviews surpassed that of the five manual reviews."[13] The results indicated that the technology-assisted reviews yielded, on-average, a higher recall than the manual reviews, although the difference was not statistically significant.[14] The precision of the technology-assisted review was, however, significantly higher, and thus more accurate than the manual review.[15] A study conducted in 2009 similarly reflected that "[o]n every measure, the performance of the two computer systems was at least as accurate (measured against the original review) as that of human re-review."[16]

Doesn't Replace Human Review

As conceived, the technology for predictive coding review will not replace human review, but rather is intended to have high-level attorneys train the automated tool in order to accelerate the e-discovery process, reduce the cost associated with exhaustive manual reviews and increase the accuracy of the overall review. Sophisticated counsel are essential to the e-discovery process. Although predictive coding leverages intelligent technology, it is no substitute for human intelligence, which is necessary to analyze the documents in relation to the theories of the case.[17] Whenever using technology, even sophisticated technology, the maxim "garbage in, garbage out" remains true.

In order to defend the use of predictive coding as a reasonable method to handle the review and production of ESI, careful consideration and hands-on management of senior attorneys and their litigation consultants is required. As mentioned above, a typical first-level document reviewer should not train a predictive coding system. The proper training of a predictive coding tool should be done by a senior attorney who has thoughtfully reviewed the document requests, analyzed the issues in the case, and considered a scope of relevance for the review. Both outside counsel and clients should coordinate their efforts to identify key documents to ensure that the system is including them when searching across the client's population of potentially responsive ESI. Litigation consultants, too, need to fully understand the predictive working tool and the protocols to be used in order to defend the process as reasonable.

An important aspect of e-discovery generally is communication with opposing counsel. While there is no legal requirement that mandates a party to proactively disclose it is using predictive coding technology, transparency and cooperation with an adversary is emphasized in the comments to the Federal Rule of Civil Procedure and is a best practice for ESI practitioners. Vendors promoting predictive coding technology have made efforts to make their tools user friendly, enabling both clients and clients' adversaries to track "how their systems [are] used to select records."[18] In addition, many systems' searches can be verified by replicating search results on the same data "if the steps outlined in the audit trail are followed."[19]

As with e-discovery practices in general, predictive coding processes should be well-documented in a discovery protocol that includes how the sample sets were identified, how the seed set was comprised, and how many iterations are necessary to be confident in the final population. Seed sets and documents used for the iterative reviews should be preserved, and parties should be open to sharing these documents (less the privileged documents) with adversaries to help achieve buy-in. Of course, if parties cannot agree on whether or how predictive coding technology shall be employed, magistrate judges, special masters and referees can resolve the differences, as Judge Peck did.

Judicial Acceptance

As indicated, on Feb. 24 Judge Peck issued an opinion in *Da Silva Moore*, "recogniz[ing] that computer-assisted review is an acceptable way to search for relevant ESI in appropriate cases."[20] In *Da Silva Moore*, defendant, MSLGroup, proposed to limit the cost of its review and production of over 3.2 million documents to $550,000 by utilizing predictive coding software.[21] When plaintiffs questioned defendants' proposed use of predictive coding technology, defendants must have been pleased the case was referred to Magistrate Judge Peck, who recently authored an article in

Law Technology News including his view that "computer-assisted coding should be used in those cases where it will help 'secure the just, speedy, and inexpensive' (Fed. R. Civ. P. 1) determination of cases in our e-discovery world."[22]

In *Da Silva Moore*, the parties attempted to negotiate a joint ESI protocol including predictive coding technology; however, they were unable to reach an agreement. Appearing before the court on Feb. 8, they presented their positions on a proposed ESI protocol and a method for the use of predictive coding, and Judge Peck generally ruled in favor of defendant, complimenting defendant on the transparency of the proposed methodology for predictive coding and reminding plaintiffs that the goal of using predictive coding is not perfection but to improve on the status quo.[23] At the conclusion of the hearing, Judge Peck ordered the parties to submit a joint protocol on Feb. 17, based upon his rulings and indicated that upon review, he may issue an opinion on the use of predictive coding "for the benefit of the greater bar."[24]

The parties were unable to reach agreement on a joint submission by Feb. 17. Thus, they submitted an unsigned "Proposed Protocol Relating to the Production of Electronically Stored Information (ESI)," which incorporated Judge Peck's orders from the Feb. 8 status conference.[25] With respect to predictive coding, the protocol details what the random sample will be,[26] how the seed set will be identified,[27] the methodology for the iterative reviews,[28] and defendant's proposal for quality control.[29] Plaintiffs included their objection to the protocol, in its entirety.[30] Nevertheless, on Feb. 22, Judge Peck so ordered the protocol, with a few modifications.[31] Plaintiffs promptly appealed the order to the district court.

Judge Peck's opinion is the first formal judicial opinion approving the use of predictive coding for handling ESI.[32] Judge Peck determined that the use of predictive coding was appropriate in *Da Silva Moore* considering:

(1) the parties' agreement,

(2) the vast amount of ESI to be reviewed (over three million documents),

(3) the superiority of computer-assisted review to the available alternatives (i.e., linear manual review or keyword searches),

(4) the need for cost effectiveness and proportionality under Rule 26(b)(2)(C), and

(5) the transparent process proposed by MSL.[33]

He also enumerated several "lessons for the future" for counsel to take-away from the resolution of the discovery disputes in this case. Judge Peck advised: First, although cost is a factor, courts will be unlikely to set a limit on document review and production until the results of the review are quality-control verified. Second, parties should consider "staging of discovery" by starting with the most likely relevant discovery sources, as a way to control costs. Third, counsel should utilize their clients' knowledge of the opposition's, as well as their own, most relevant custodians and sources of ESI. And fourth, e-discovery vendors have a helpful role to play at court conferences to explain complicated ESI concepts in a way that is "easily understandable to judges who may not be tech-savvy."[34] His main take-away though, is that predictive coding is an available tool that can and should be effectively used to save parties time and money in discovery.[35]

••••••••••••●●●••••••••••••••

[1] The term "predictive coding" does not describe or endorse any vendor or any vendor's product. "Predictive coding" is also commonly referred to as "computer-assisted review" or "technology-assisted review."

[2] Anne Kershaw & Joseph Howie, "Crash or Soar? Will the legal community accept 'predictive coding?'" Law Technology News (Online), Oct. 1, 2010, available at http://www.akershaw.com/articles/LTN_CrashOrSoar_2010_Oct.pdf.

[3] Transcript of Feb. 8, 2012 Status Conference at 93, *Da Silva Moore v. Publicis Groupe*, 11 Civ. 1279 (S.D.N.Y. Feb. 8, 2012) (No. 88) (hereinafter Transcript).

[4] *Da Silva Moore v. Publicis Groupe*, 11 Civ. 1279 (S.D.N.Y. Feb. 24, 2012) (No. 96) (Opinion & Order approving the use of predictive coding) (*Da Silva Moore*).

[5] Id. at 25.

[6] Id. at 26.

[7] David Blair and M.E. Maron, "An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval System," 28 Comm. ACM 289, 290 (1985).

[8] Kershaw & Howie, supra note 2.

[9] Id.

[10] Id.

[11] Maura R. Grossman & Gordon V. Cormack, "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review," 17 Rich. J.L. & Tech. 11, Spring 2011, at 35, available at http://jolt.richmond.edu/v17i3/article11.pdf.

[12] Id. at 35-37.

[13] Id. at 43.

[14] Id. at 43-44.

[15] Id.

[16] Herbert L. Roitblat, et al., "Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review," 61 J. Am. Soc'y for Info. Sci. and Tech. 70, 79 (2010).

[17] Caroline Privault, et al., "A New Tangible User Interface for Machine Learning Document Review," 18 Artif. Intell. Law 459, 460 (2010).

[18] Kershaw & Howie, supra note 2.

[19] Id.

[20] *Da Silva Moore v. Publicis Groupe*, 11 Civ. 1279 (S.D.N.Y. Feb. 24, 2012) (No. 96), at 2.

[21] *Da Silva Moore v. Publicis Groupe*, 11 Civ. 1279 (S.D.N.Y. Feb. 25, 2012) (No. 92) (Parties' Proposed Protocol Relating to the Production of Electronically Stored Information & Order).

[22] Andrew J. Peck, "Search, Forward," Law Technology News (Online), Oct. 1, 2011, available at http://www.law.com/jsp/lawtechnologynews/PubArticleLTN.jsp?id=1202516530534&slreturn=1.

[23] Transcript, supra note 3, at 57-94.

[24] Id. at 93-94.

[25] ESI Protocol & Order, supra note 21.

[26] Id. at 13.

[27] Id. at 14-15.

[28] Id. at 15-17.

[29] Id. at 17-18.

[30] Id. at 22.

[31] Id. at 1; see also id. at 17, 18 (rejecting defendant's proposed provisions re: cost-shifting).

[32] *Da Silva Moore*, supra note 4 at 25.

[33] Id. at 22.

[34] Id. at 23-25.

[35] Id. at 26.

Reprinted with permission from the March 19, 2012 edition of the NEW YORK LAW JOURNAL © 2012 ALM Media Properties, LLC. All rights reserved. Further duplication without permission is prohibited. For information, contact 877-257-3382 or reprints@alm.com. # 070-03-12-29