Predictions Regarding Predictive Coding

Law360, New York (October 31, 2012, 4:49 PM ET) -- Although the technologies involved with the concept of "predictive coding" have emerged over the last several years and are far from enjoying widespread acceptance, recent developments may signal a new chapter in electronic discovery in civil litigation. As discovery has grown to encompass electronic documents, technology has evolved, culminating most recently with the concept of technology-assisted document review — so-called predictive coding. Indeed, the first judicial opinion regarding predictive coding was issued just this year.

What is Predictive Coding?

The phrase "predictive coding" may not familiar to many, but analogous concepts are actually becoming fixtures of our daily lives through technologies such as those used by Pandora and Netflix.[1] Predictive coding utilizes "sophisticated algorithms ... [which] enable the computer to determine [the] relevance [of documents], based on interaction with (i.e., training by) a human reviewer."[2]

It relies on a human to code a sample of the documents, called a "seed set," that in turn allows the computer to identify properties of those documents and then evaluate the remaining documents, looking for similar characteristics. Generally speaking, the goal of predictive coding is to efficiently produce a set of documents that has high recall — meaning that all of the responsive documents are included — and precision — meaning that only the responsive documents are included.

Predictive coding is accomplished through a series of steps. First, the seed set is selected. This is an important step that requires an understanding of both the facts of the case and the technical properties of the computer software. The seed set must provide the software with the right kind of information to allow the algorithm to process the remaining documents. For instance, if the algorithm works by identifying linguistic patterns in relevant documents, a seed set filled with irrelevant documents will be ineffective.

Next, a lawyer with a strong grasp of the subject matter of the case will review and code the seed set for relevance, privilege and, in some cases, the issues to which each document relates. Finally, the computer software's design will determine the next steps. In most cases, after the seed set is coded, the computer generates additional sets that are again coded by the lawyers on the case. The goal is that after several iterations, the computer software will reach a point where its predictions are accurate to a statistically significant level, meaning that the computer is able to consistently predict the proper coding for the rest of the documents in the set.

Guidance from the Courts

Da Silva Moore v. Publicis Groupe, No. 11 Civ. 1279 (ALC)(AJP) (S.D.N.Y. Feb. 24, 2012)

Not surprisingly, since predictive coding is still a nascent area of the law, relatively few

cases have been decided. In fact, several months before the Da Silva case was assigned to Magistrate Judge James Peck, he published an article about predictive coding noting that it appeared that many lawyers were waiting for a judicial decision approving the use of predictive coding. Da Silva has been described by some as a landmark case on predictive coding. In Da Silva, the parties had agreed to implement the technology. What remained were the parties' disputes regarding the defendant's proposed protocol and production requirements.

While Magistrate Judge Peck suggested that predictive coding should only be used in appropriate circumstances, he also described the drawbacks of using keyword searching as the sole review technology.[3] Namely, keywords can be ineffective and parties often use them to conduct fishing expeditions. Then, Magistrate Judge Peck discussed why predictive coding was appropriate in Da Silva, emphasizing the importance of the Federal Rules of Civil Procedure and their mandate to "secure the just, speedy, and inexpensive" resolution of the case.[4] He also invoked what he viewed as the proportionality doctrine of Rule 26(b)(2)(C).

His results-driven standard recognized that the extent of measures taken must be proportional to the costs, amount in controversy and the quality of the predictive coding output.[5] Consequently, predictive coding was appropriate in light of (1) the parties' agreement, (2) the volume of electronically stored information to be reviewed (over 3 million documents), (3) the superiority of computer-assisted review to the available alternatives, (4) the need for cost effectiveness and proportionality under Rule 26(b)(2)(C), and (5) the transparent process proposed by the defendant.[6]

The transparency consideration is particularly interesting. In the protocol approved by Magistrate Judge Peck, the seed set comprised several thousand documents generated through a three-step process using random sampling and keyword searches proposed by both parties. The defendant agreed to produce the entire seed set, except for documents protected by privilege, so the plaintiffs could review the documents (both responsive and nonresponsive) and their coding.

After coding the seed set, defendant's counsel would conduct seven rounds of iterative review to "stabilize" the software. In each successive round, a human coder would review a set of 500 documents generated by the computer and either confirm or correct the software's predictions. Finally, the defendant's counsel would draw and review a random sample from the null set (i.e., the documents the software coded as non-relevant), with the aim of confirming high levels of precision and recall.

The court held that if after the seven rounds the software had not reached a sufficient level of stabilization, the defendant would have to "do another round or two or five or 500 or whatever it takes to stabilize the system."[7] However, under principles of proportionality, the court suggested that if the null set included relevant documents that were "more of the same" and "d[id] not add anything to the case," it might not matter.[8] On the other hand, if so-called "smoking gun" or "hot" documents remained in the null set, more training of the software might be necessary.[9]

Although the plaintiffs' counsel expressed concerns regarding the reliability of predictive coding and urged the court to adopt ex ante standards, Magistrate Judge Peck disagreed, concluding that the only way to determine whether the protocol was effective was to review the results ex post. The court was not concerned with potential ambiguities in the standards for relevance because the proposed level of transparency would allow the plaintiffs to review the coding and raise issues with the court as needed.

In April 2012, Judge Robert Carter adopted Magistrate Judge Peck's opinion, finding it critical that the "standards for measuring the reliability of the process and the protocol builds in levels of participation by [p]laintiffs. It provides that the search methods will be carefully crafted and tested for quality assurance, with [p]laintiffs participating in their implementation."[10]

Citing the principles of Rules 1 and 26(b)(2)(C) of the Federal Rules of Civil Procedure, Judge Carter further noted that "there is simply no review tool that guarantees perfection."[11] Since the magistrate concluded that predictive coding was more appropriate than keyword searches and he implemented a mechanism to resolve future disputes between the parties, Judge Carter adopted Magistrate Judge Peck's ruling.[12]

Global AeroSpace v. Landow Aviation LP, No. CL 61040 (Va. Cir. Ct. April 23, 2012)

In Landow Aviation, the defendants asked the court to either approve of their use of predictive coding or order the plaintiffs to bear any added costs of human review. The defendants urged the court to approve their proposed predictive coding protocol because a large volume of documents would need to be reviewed and a first pass human review would require approximately 20,000 man-hours — at an estimated cost of $2 million and with a projected recall rate of 60 percent.[13]

The defendants argued that, in their case, the benefits of predictive coding outweighed the drawbacks in terms of time savings and recall and precision.[14] The plaintiffs disputed that the software could reach a sufficient level of accuracy and also declined to participate in the process, since they saw the production as too small to necessitate going through an iterative process.

Landow Aviation is significant in part because the court authorized the defendants to utilize predictive coding over the plaintiffs' opposition .[15] However, Judge James Chamblin took a slightly different approach from Magistrate Judge Peck's in Da Silva. Rather than analyzing the technical advantages predictive coding may have over human review, Judge Chamblin analogized the issue to controversies that arose regarding previous technologies. He noted that when large document productions first became common, parties would argue about whether junior associates should be allowed to review documents instead of more senior lawyers.

In the court's view, that debate is similar to the current debate, because both involve decisions regarding how to best address the practicality of reviewing large electronic document sets.[16] The court thought predictive coding was worth trying, because it offered potential cost savings and quality that were at least as good as what could be expected from human review. Ultimately the court allowed the defendants to explore the option, subject to the plaintiffs' right to raise concerns or objections with the court.[17]

Kleen Products LLC v. Packaging Corp. of America, No. 1:10-cv-05711 (N.D. Ill. Aug. 21, 2012)

In Kleen Products, the plaintiffs sought to force predictive coding on the defendants even though the defendants had already produced documents using keyword searches.[18] The court conducted two days of hearings, which included testimony by expert witnesses in the area of e-discovery. Ultimately, rather than opine on the merits of one technology over another, Magistrate Judge Nan Nolan emphasized the need for the parties to cooperate in the keyword searching in order to allow the case to progress.

Magistrate Judge Nolan relied on Sedona Principle 6, which states that "[r]esponding parties are best situated to evaluate the procedures, methodologies, and technologies appropriate for preserving and producing their own electronically stored information."[19] She asked the parties to agree upon a set of keyword search protocols instead of battling about which technology would yield better results.[20] The parties ultimately reached a stipulation by which the plaintiffs withdrew their initial challenge to defendants' use of keyword searching, but reserved their rights to object to the details of the defendants' keyword searching methodology and to demand that predictive coding be used in later requests.[21]

In re Actos (Pioglitazone) Products Liability Litigation, No. 6:11-MD-02299 (W.D. La. July 27, 2012)

In In re Actos, the parties already had stipulated to a predictive coding protocol, and the court incorporated the parties' agreement in the Case Management Order.[22] The order included a "Search Methodology Proof of Concept," which described the parties' agreement that after the privileged documents were extracted, both parties would have the opportunity to code all of the documents in the seed set, responsive or not. Likewise, the order dictated that the parties would meet and confer regarding any conflicting coding decisions.[23]

In re Actos appears to have followed the Da Silva model by forcing the parties to make joint decisions about relevance in training the software. Thus, In re Actos extends the theme of cooperation seen in the earlier cases. By allowing for joint relevance review, the parties have a say on the front end. In issuing the order, the court showed faith in the new technology and the safeguards that were proposed to ensure its reliability.

EORHB Inc. v. HOA Holdings LLC, No. 7409-VCL (Del. Ch. Oct. 15, 2012)

Most recently, Vice Chancellor Travis Laster of the Delaware Court of Chancery endorsed the use of predictive coding in the context of a non-expedited indemnification proceeding. He ordered the parties to show cause if they did not want to use it, and also ordered that the parties agree on a single discovery provider that could warehouse both parties' documents and "maintain the integrity of both side's [sic] documents and insure [sic] that no one can access the other side's information."[24] He found these proceedings to be an "ideal" context for predictive coding because "these types of indemnification claims can generate a huge amount of documents."

His view was that it was better to employ technology-assisted review instead of "burning lots of hours with people reviewing."[25] Vice Chancellor Laster's order from the bench illustrates the growing view that predictive coding can be an attractive tool in the appropriate circumstances. Here, where the parties had time to carefully coordinate with each other, Vice Chancellor Laster recognized the opportunity to implement the cutting-edge discovery technology.

Practical Takeaways

Across these cases, two themes emerge. First, Rules 1 and 26(b)(2)(C)(iii) of the Federal Rules of Civil Procedure are prominent in these cases. Rule 1 states that procedural decisions should be made "to secure the just, speedy, and inexpensive determination of every action and proceeding," while Rule 26(b)(2)(C)(iii) provides that a court can limit discovery if "the burden or expense of the proposed discovery outweighs its likely benefit, considering the needs of the case, the amount in controversy, the parties' resources, the importance of the issues at stake in the action, and the importance of the discovery in resolving the issues."

While not all of the judges addressed the virtues of one technology over another, all emphasized and approved of parties' cooperation during the discovery process. Different cases may require the use of varying discovery tools, but adherence to the principles of the Federal Rules is critical to effective advocacy.

Second, these decisions indicate that, at present, predictive coding remains one option that should be considered, based on the particular circumstances of each case. Magistrate Judge Nolan asked the parties to cooperate so that proceedings on other pressing issues could take place. Meanwhile, in Da Silva, Landow Aviation, In re Actos and EORHB, the parties and the courts all committed to testing the waters with predictive coding, subject to the potential for additional hearings on disputed issues. These cases make clear that the standards being applied to predictive coding issues are not uniform and, at present, are best resolved on a case-by-case basis.

Predictive coding is not, at this point, a replacement for predecessor technology, nor is it a substitute for human review. Rather, all available technologies should be seen as tools available to practitioners that might be useful when deployed under the proper

circumstances. In fact, predictive coding may be useful to practitioners in contexts beyond the bounds of document production that require the review of large sets of electronically stored information (e.g. in witness preparation or summary judgment briefing).

It is also worth noting that selecting the right e-discovery vendor to assist with predictive coding could be crucial. With widely divergent technologies, price structures, products and services, it will be important for attorneys to retain and work with trustworthy and well-reputed e-discovery vendors, who can help to ensure that a chosen technology is appropriate in a given case. For instance, predictive coding may not be the best choice for document sets that consist primarily of non-email or text-based documents (such as Microsoft Word files and .txt files). Documents with graphics, numbers or text that cannot be subjected to optical character recognition technology may not be as effectively reviewed by computer software.

Although it is just beginning, the trend toward new technologies in predictive coding may be picking up speed. In a recent opinion in National Day Laborer Organizing Network v. United States Immigration & Customs Enforcement Agency, Judge Shira Scheindlin questioned the continued viability of relying solely on keyword searching. [26] She explained that "beyond the use of keyword search, parties can (and frequently should) rely on latent semantic indexing, statistical probability models, and machine learning tools to find responsive documents. Through iterative learning, these methods (known as "computer-assisted" or "predictive" coding) allow humans to teach computers what documents are and are not responsive ..."[27]

--By Lauren Aguiar and Jonathan Friedman, Skadden Arps Slate Meagher & Flom LLP

Lauren Aguiar is a partner at Skadden in New York. Jonathan Friedman, an associate in the firm's New York office, provided assistance with the article.

The opinions expressed are those of the authors and do not necessarily reflect the views of the firm, its clients, or Portfolio Media Inc., or any of its or their respective affiliates. This article is for general information purposes and is not intended to be and should not be taken as legal advice.

[1] It does not appear that predictive coding has entirely replaced (or will replace) familiar predecessor technology such as keyword searching, as keyword searching of electronic data is still used by many practitioners both as a primary method of review optimization or as one aspect of other protocols.

[2] Andrew Peck, Search, Forward, L. Tech. News, Oct. 1, 2011, available at http://www.law.com/jsp/lawtechnologynews/PubArticleLTN.jsp?id=1202516530534.

[3] Da Silva, 2012 WL 607412, at *3 (but approving of the parties' use of Boolean searching in generating the seed set as a part of their predictive coding protocol).

[4] Id. at *1 (quoting Fed. R. Civ. P. 1).

[5] Id. at *12.

[6] Id. at *11.

[7] Id.

[8] Da Silva, 2012 WL 607412, at *8.

[9] Id.

[10] Da Silva Moore v. Publicis Groupe, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 1446534, at *2 (S.D.N.Y. Apr. 26, 2012).

[11] Id.

[12] Id. at *3.

[13] See Memorandum in Support of Motion for Protective Order Approving the Use of Predictive Coding, Global AeroSpace, Inc. v. Landow Aviation, L.P. (Va. Cir. Ct. filed Apr. 9, 2012) (No. CL 61040), 2012 WL 1419842.

[14] See Transcript of Oral Arguments on Motions on Behalf of Defendants, Counter-Defendants, Cross-Defendants and Third Party Defendants at 143-46, Global AeroSpace v. Landow Aviation, L.P. (No. CL 61040) (Va. Cir. Ct. filed Apr. 23, 2012).

[15] See Order Approving the Use of Predictive Coding for Discovery, Global AeroSpace, Inc. v. Landow Aviation, L.P. (No. CL 61040) (Va. Cir. Ct. filed Apr. 23, 2012), 2012 WL 1431215.

[16] Transcript of Oral Arguments, supra note 15, at 172.

[17] Id. at 173-74.

[18] Plaintiffs' Reply Memorandum of Law for Evidentiary Hearing at 6-8, Kleen Products LLC v. Packaging Corp. of America (No. 10-cv-05711) (N.D. Ill. filed Feb. 13, 2012).

[19] Transcript of Proceedings at 297-98, Kleen Products LLC v. Packaging Corp. of America (No. 10-cv-05711) (N.D. Ill. Mar. 28, 2012) (emphasis added).

[20] Id. at 297-300 (noting that the parties had indicated they could spend the next two years arguing motions to compel).

[21] See Stipulation and Order Relating to ESI Search at 3, Kleen Products LLC v. Packaging Corp. of America (No. 10-cv-05711) (N.D. Ill. filed Aug. 21, 2012).

[22] Case Management Order: Protocol Relating to the Production of Electronically Stored Information ("ESI"), In re Actos (Pioglitazone) Products Liability Litigation (No. 11-MD-02299) (W.D. La. filed Jul. 27, 2012).

[23] Id. at 8.

[24] Transcript of Proceedings at 66-67, EORHB, Inc. v. HOA Holdings LLC, No. 7409-VCL (Del. Ch. Oct. 15, 2012).

[25] Id.

[26] National Day Laborer Organizing Network v. United States Immigration & Customs Enforcement Agency, No. 10 Civ. 3488 (SAS), 2012 WL 2878130, at *12 (S.D.N.Y. July 13, 2012) (to be published in F. Supp. 2d).

[27] Id.