Westlaw Journal
COMPUTER & INTERNET
Litigation News and Analysis • Legislation • Regulation • Expert Commentary
VOLUME 32, ISSUE 6 / AUGUST 28, 2014

EXPERT ANALYSIS

Predictive Coding 2.0

By Joe White, Esq., *Kroll Ontrack*, and
Cliff Nichols, Esq., *Day Pitney LLP*

Predictive coding is finally catching on.[1] While we have discussed for years its ability to dramatically decrease costs over standard linear review, there is a growing consensus that recognizes it as also performing more efficiently and accurately. Predictive coding can help cull through mountains of data, quickly identifying relevant and responsive documents, making the document review process much less time consuming for practitioners with a more accurate and precise result than previously accepted discovery search methods.

But this powerful technology should not be limited to discovery alone, as it is capable of much, much more. Imagine predictive coding being used to identify potential risks to a company before these risks develop into litigation. What if predictive coding could organize incoming e-mails in order of priority? What about a real-time review of outgoing documents that identifies potential leaks of trade secrets?

The idea of incorporating machine learning and predictive coding into everyday business tasks is not far-fetched. In fact, many retail companies have already implemented basic forms of the technology to create tools such as automated wish lists and product recommendations. The potential scope of application may be limitless.

WHY PREDICTIVE CODING?

In the discovery context, predictive coding is rapidly replacing keyword searches and other traditional search methods as it is accepted by judges and courts as a valid discovery method.[2] In fact, some courts have gone so far to say that keyword searching is "usually not very effective" and parties should use it in appropriate cases.[3] In effect, this goes further than signaling its acceptance by courts; it also implies that practitioners are well-advised to leverage the technology in suitable cases or lose out.

This judicial acceptance should also help pave the way for an expanded use of predictive coding outside of discovery. In-house counsel and compliance officers, invariably both cautious groups, can point to the success of predictive coding and machine learning in discovery when determining how it can be effective in other areas.

So what is it that makes predictive coding so effective as opposed to other data search and filter methods? The answer to that question lies in the workflow of the predictive coding process and the inherent inaccuracy of other methods. Keyword searching can limit the number of documents reviewed but you are limited by your, often uninformed, imagination. Keyword searching only really helps you find documents that you already know, or suspect, exist. To the extent you find anything else, you stumble across it.

Keyword searches can also produce over-broad or under-broad data sets. If too broad, it can still be a lot of work to review every document and pick out those that are relevant. If under-broad, the search may not produce all the relevant documents as it was limited by the contemplation of the searcher.

Though still a helpful tool, the major drawback of keyword searching is that it can be both over-broad and under-broad, *at the same time*. It's easy to envision a scenario where, in a gender discrimination suit, a keyword search of "female" may be so broad that it is a major task to review the responsive documents, while also being so narrow that it does not consider possibly relevant documents that do not use that descriptor. As U.S. Magistrate Judge Andrew J. Peck of the Southern District of New York put it, "In too many cases ... the way lawyers choose keywords is the equivalent of the child's game of 'Go Fish.'"[4]

This is one of the reasons that predictive coding is evolving as a preferred method of discovery search. What makes predictive coding exceptional is that you do not necessarily need to know what documents exist to find what you need. Predictive coding learns what you want, often before you do, and uses a scientifically verifiable way to find them.

HOW IT WORKS

Simply put, predictive coding works by having human experts manually "categorize" a subset of documents, which the technology then uses to create a predictive model. This model is applied to the larger data set and can be used to estimate the probability that a document belongs to each of the possible categories. Predictive coding can utilize a document's content features, such as words and phrases, and its metadata features and decide how much weight to give each of those features. Therefore, it is not limited to exact matches in the same way as keyword searches. Not only does predictive coding help find which haystack the proverbial needle may be in, it also reduces the amount of hay in that stack.

Predictive coding can help cull through mountains of data, quickly identifying relevant and responsive documents and making the document review process much less time consuming for practitioners.

An example of simple predictive coding in everyday life is the music website or mobile app Pandora. When you tell Pandora that you want to listen to "Led Zeppelin," for example, you have created a "seed set" that teaches Pandora that you want songs recorded by Led Zeppelin and that you may also like other British hard rock from the 1970s with strong guitar and sometimes-mystical themes. Pandora's software will automatically search for music that has some association to these characteristics.

In addition to Led Zeppelin music, Pandora may also suggest Deep Purple, Rainbow or Cream. By using the "thumbs up" or "thumbs down," you further train the program on what kind of music you are specifically looking for and the subset becomes smaller and more accurate. Before long, and with proper training, you'll never again have to listen to that one-time folk song Justin Timberlake covered. But, you may have to hear Nena's "99 Luftballons," so nothing is perfect.

WHAT'S NEXT?

Many potential uses for predictive coding exist,[5] and we discuss below three developing use cases where organizations and practitioners may use it to create value.

Early data assessment

While some practitioners may have familiarity in using predictive coding in the later stages of discovery, many may not have recognized its value in the early stages of a case before the data to be searched is even identified. Early data assessment, or EDA, ideally narrows the scope of potential important data, and gives attorneys information about the applicability of documents, early in the case. This is done by separating data into critical and non-critical groupings, and identifying and narrowing the number of actual key players and search terms. This usually involves a process of searching, foldering, clustering, email threading and topic grouping. Unfortunately, predictive coding is not often implemented in these early stages, but it could produce immense value.

In addition to narrowing the scope of collection and refining the precision of collection, early use of predictive coding can be used to confidently effect settlement. In *New Mexico State Investment Council v. Bland*, the court recognized the value of using predictive coding in EDA.[6] The court also noted that this approach avoided "prolonged and extensive discovery" and helped effect

early settlement.[7] The pre-discovery investigation involved multiple databases and more than 5 million documents. Using predictive coding, the investigative team was able to narrow the responsive documents down to under 50,000. The result was a much smaller human review.

Compliance tool

Stopping the genie before it gets out of the bottle can be difficult, but early detection of potential problems can reap great dividends. The problem is the difficulty of predicting bad behavior before it is too late. Predictive coding is a natural way to assess and detect risk patterns, and stop them from developing further.

As explained above, if trained properly, software can use predictive coding to identify any information it has been taught to detect as relevant. Therefore, predictive coding software could be trained to comb through all an organization's information to detect potential risks. These risk-signaling documents could then be grouped and reviewed to determine if a real risk to the organization exists. In this way, counsel could quickly take immediate corrective action, ideally preventing lawsuits before they occur.

The biggest challenge is training the computer program to identify the appropriate risks. No two cases are alike, but they often contain similar patterns. An organization need only identify what frequent, undesirable patterns exist in recurrent litigation and train the software to recognize these patterns.

Consider the recent litigation of *Securities and Exchange Commission v. Biomet Inc.*[8] In that case, alleged violations of the Foreign Corrupt Practices Act cost Biomet $22 million to settle. The SEC alleged that Biomet paid bribes to foreign, publicly employed doctors in violation of the FCPA. Allegedly, Biomet employees recorded these bribes as "consulting fees" or "commissions" using phony invoices. The SEC claimed that Biomet sent emails describing bribery in no uncertain terms.

With that knowledge, predictive coding could be used to train an email system to scan for patterns based on this type of activity. The software could detect a pattern of recorded consulting fees being paid to foreign officials and flag the records for further review, which could possibly save Biomet or other organizations millions the next time around. What makes predictive coding uniquely qualified from a compliance standpoint is that the technology can be trained to detect the emails that describe bribery, even if the word itself is never explicitly used.

This is one example, and it is easy to imagine others where predictively auditing an organization's documents in real time could reveal undesirable trends. Getting conservative and often frugal boards to see the "long game" and actually implement the technology may be difficult, however. The "bird in the hand is worth two in the bush" perspective pervades business often with very real consequences.

Further, if predictive coding is used to "sweep" internal systems, and potential litigation risks are identified, this could trigger the duty to preserve data, as a court may rule that litigation is reasonably anticipated at the point of identification. Still, it is almost always better to identify potential problems early rather than be caught unaware later. The "head in the sand" approach does not play well with judges.[9] In addition, operating a highly functional compliance program with systems that effectively identify lapses can positively impact the course of a government investigation in a significant way.

Defensible record retention policies

Finally, predictive coding can be and should be used to enforce and create record retention policies. We can use predictive coding to identify documents that should be kept — and more importantly which ones should not — so the glut of extraneous data can be defensibly decreased. Because of sheer volume, a more automated approach to records retention is the undeniable future of information governance.

As many in legal and IT departments of organizations know, record retention policies are only good if employees follow them. Even if employees did their best to abide by a record retention

The idea of incorporating machine learning and predictive coding into everyday business tasks is not far-fetched.

policy, there is always a risk of human error as documents can easily be mis-categorized or deemed irrelevant when they are anything but. By amplifying quality training inputs with the speed and consistency of machine-supplied outputs, predictive coding can do much to reduce this risk by correctly identifying documents that must be kept.

Given the speed and volume at which data is created, it's not hard to imagine that the data stored on any organization's servers might become disorganized and chaotic. Consider a scenario in which documents containing numeric values often considered confidential (such as Social Security numbers, credit card information, phone numbers, etc.), are improperly stored in a non-secure location on company servers. Predictive coding could be used to group and categorize this data so it could quickly be determined whether it is ripe for deletion or if it should be moved to a more secure location. This use works best when coupled with the internal compliance and risk assessment features given in the previous example. In this way, businesses can build or enforce a record retention and information governance policy that is predictable, organized and defensible.

By organizing the data, and weeding out all unnecessary documents, businesses would be better equipped to more cost-effectively scan for potential litigation risks, as well as produce for the discovery stage of litigation should a suit be filed. The drawback of using predictive coding for records management and defensible deletion is the very reason many businesses choose not to delete any information at all: the risk of deleting something that may be discoverable in ongoing litigation. Such an approach, however, is unsustainable. Even if all the data were retained, this just puts off the inevitable, meanwhile spending more money and subjecting an organization to more risk later on when the data needs to be reviewed for litigation.

In the discovery context, predictive coding is rapidly replacing keyword searches and other traditional search methods as it is accepted by judges and courts as a valid discovery method.

CONCLUSION

Predictive coding is a powerful technology that is currently not being used to its full potential. Organizations could use it for much more than merely responding to discovery requests. If its full capability was leveraged, there are countless use cases where the technology can drive value for organizations. EDA, risk assessment and records management are simply three examples of predictive coding's potential. The Nobel laureate Niels Bohr has been quoted as saying, "Prediction is very difficult, especially about the future,"[10] but with predictive coding, it is not quite as difficult as it used to be.

NOTES

[1] *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182, 193 (S.D.N.Y. 2012) ("Counsel no longer have to worry about being the 'first' or 'guinea pig' for judicial acceptance of computer-assisted review."), *adopted sub nom. Moore v. Publicis Groupe SA*, 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012).

[2] *Id.* at 191 ("Computer-assisted review appears to be better than the available alternatives, and thus should be used in appropriate cases."); *Nat'l Day Laborer Org. Net. v. U.S. Immigration & Customs Enforcement Agency*, 877 F. Supp. 2d 87, 109 (S.D.N.Y. 2012) ("[P]arties can (and frequently should) rely on ... [predictive coding]."); *Global Aerospace Inc. et al. v. Landow Aviation et al.*, No. CL 61040, 2012 WL 1431215 (Va. Cir. Ct., Loudoun County Apr. 23, 2012) ("[I]t is hereby ordered defendants shall be allowed to proceed with the use of predictive coding for purposes of the processing and production of electronically stored information."). For analysis on how predictive coding works in a real world example, see generally *In re Actos (Pioglitazone) Prods. Liab. Litig.*, No. 6-11-MD-2299, 2012 WL 7861249 (W.D. La. July 27, 2012).

[3] *Da Silva Moore*, 287 F.R.D. at 191. *See also Nat'l Day Laborer*, 877 F. Supp. 2d at 109 ("Simple keyword searching is often not enough.").

[4] *Id.* at 190-191 (citing Ralph C. Losey, *Child's Game of 'Go Fish' is a Poor Model for e-Discovery Search*, in ADVENTURES IN ELECTRONIC DISCOVERY 209–10 (2011)).

[5] *See, e.g.,* Brad Blickstein, *Predictive Coding: It's Not Just For Discovery Anymore*, INSIDE COUNSEL MAGAZINE, May 22, 2014, http://www.insidecounsel.com/2014/05/22/predictive-coding-its-not-just-for-discovery-anymo, for a discussion of how predictive coding can impact legal billing.

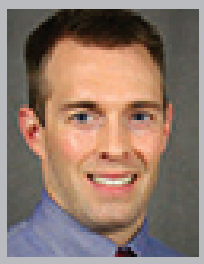
[6] 2014 WL 772860, at *5 (N.M. Dist. Ct. Feb. 12, 2014) ("In reviewing documents, Day Pitney implemented various advanced machine learning tools such as predictive coding, concept grouping, near-duplication detection and e-mail threading. .... [This] enabled the reviewers on the document analysis teams to work more efficiently with the documents and identify potentially relevant information with greater accuracy than standard linear review.").

[7] *Id.* at *17.

[8] No. 1:12-CV-00454, *complaint filed* (D.D.C. Mar. 26, 2012).

[9] *Gonzalez-Servin v. Ford Motor Co.*, 662 F.3d 931, 934 (7th Cir. 2011) ("The ostrich is a noble animal, but not a proper model for an appellate advocate.").

[10] Felecity Pors, *The Perils of Prediction, June 2nd,* THE ECONOMIST, July 15, 2007, http://www.economist.com/blogs/theinbox/2007/07/the_perils_of_prediction_june.



Joe White (L) is a senior discovery services consultant for **Kroll Ontrack** in Eden Prairie, Minn., where he focuses on education, collaboration, support and product development for predictive coding and advanced technology intended to streamline and improve the process of finding and organizing relevant information. He can be reached at JWhite@KrollOntrack.com. **Cliff Nichols** (R) is e-discovery counsel and directs all electronic investigation at **Day Pitney LLP** in Stamford, Conn. He has significant and extensive experience representing financial institutions in connection with insider trading, fraud, and Securities and Exchange Commission compliance. He is recognized as a leader in the cost-saving and efficient use of predictive coding and other types of technology-assisted review. He can be reached at cenichols@daypitney.com.