



THINKING OUTSIDE THE BOX: USE PREDICTIVE CODING AS A RIM TOOL

Doug Smith

Predictive coding is causing quite a stir, receiving increased attention in the past year due in large part to two cases being litigated in New York and Illinois. (See the sidebar “Predictive Coding in the News.”) But, what is predictive coding – and is it really a new technology?

Sharon D. Nelson, Esq., on her “Ride the Lightning” blog, offers a comprehensive definition of predictive coding that helps make it clear that it is not new. She says *predictive coding* is a “combination of technologies and processes in which decisions pertaining to the responsiveness of records gathered or preserved for potential production purposes ... are made by having reviewers examine a subset of the collection and having the decisions on those documents propagated to the rest of the collection without reviewers examining each record.”

E-discovery systems have been using processes like this for many years. Predictive coding was developed because the proliferation of electronically stored information (ESI) being created and stored made it extremely difficult and expensive to locate relevant information that needed to be preserved and produced for litigation and investigations. So, while there have been new developments in the technology, predictive coding is not a new process.

Explaining Predictive Coding

Most predictive coding processes operate in one of two fundamental ways – either through sampling or observation.

Sampling is done by computer software, which randomly selects a subset of electronic records from all those available, presents it to a human coder for review, monitors the coder’s decisions, notes the characteristics of the records that are coded (e.g., date, recipients, author, subject,

and keywords), and then uses these recorded decisions to predict the value of the remaining documents in the collection.

In the *observation process*, the coding system monitors human coders’ actual decisions as they review records, begins to predict how a record will be coded before presenting it for coding, and then compares the predicted coding to the actual coding. Eventually, the system’s predictive coding will reach the level of accuracy that was predetermined to be sufficient. At this point, the system can be used to predict the coding decisions automatically.

Neither sampling nor the observation process relies on the computer to know anything. Each uses human decisions as a calibrating mechanism to learn about the coding details, and each could be used by an organization to classify information in use.

Using Predictive Coding for Classification

Predictive coding provides the ability to perform tasks that historically had to be done by individuals with subject matter expertise. Although the technology was developed in the e-discovery space, increasingly, organizations are discovering that it holds promise for a number of purposes. One of those purposes is information classification.

When an organization’s records are maintained using the Generally Accepted Recordkeeping Principles®, the records needed to demonstrate compliance or maintain continuity of business operations have been retained and are secure, available, and authentic. Unneeded records have been appropriately and defensibly disposed of in the normal course of business.

Maybe, though, an organization’s records are not in perfect compliance with the standards above. It is possible that people have been stashing electronic information on local drives.

And, probably not all of the information in that collaborative environment IT rolled out without the records department’s involvement is appropriately classified.

Most likely, even the document management system that is so diligently used is not as accurate as it should be since users are not perfect or uniform in making classification decisions.

Predictive coding can alleviate this situation. Using knowledge an organization already has in its systems, this technology can create classification schema for identifying and categorizing data currently held in unstructured or less formal systems.

Predictive coding is a wonderful tool in the records professional’s arsenal; it can help put an organization’s “house” in even better order than planning and training alone are able to achieve. Predictive coding also can identify areas of conflict in materials that have already been classified and provide uniformity going forward.

Begin with a Seed Set of Information

The first step is to determine what types of information the organization has that is classified in accordance with the organization’s records taxonomy. The technology underlying predictive coding can use this information as a seed set for classifying or verifying the accuracy of the organization’s classifications.

Leverage Context Provided by Data Sources

Just as the information within an organization’s financial system is classified in such a way that it can be used to provide context to other information, the document management system can be utilized in a similar fashion.

For example, any database that manages contacts for sales and marketing information will yield detailed data linking certain names to certain

lines of business or projects. And, human resources information can be used to identify when certain personnel were working in different capacities.

All of this data, when combined by the analytical tool used for predictive coding, can provide predictively classified documents for the records staff to verify. The classification system will watch the decisions of the records staff to fine tune the classification decision-making already developed.

Use the Classification Tool

In this scenario, the two models of predictive coding described at the beginning of this article were combined to provide a finely tuned classification tool. Once a classification scheme has been developed and tested using the records department's affirmations or corrections to coding, the classification system is ready to go.

With the system primed with intelligence, the records professional can now turn it loose on the data stores within the organization. Let it filter those SharePoint® pages. Depending on the complexity of the network, it may be able to access local drives, thumb drives when connected, or even remote computers when connected to the network.

Before long, information will be classified with a higher, though not likely perfect, accuracy rate.

The Bottom Line

Information needs to be classified correctly to be appropriately managed throughout its life cycle. Manual classification systems offer too many inefficiencies and inaccuracies to be used without attempting to identify more reliable systems.

Predictive coding offers a developing alternative to the manual, subjective process of coding and quality review. This technology, while currently in use as an early case assess-

Predictive Coding in the News

As courts become more familiar with predictive coding, some are explicitly endorsing and recommending the practice, lending additional credibility to the use of the technology. Three recent cases are highlighted here.

In the Circuit Court of Loudoun County Virginia, *Global Aerospace, Inc. v. Landow Aviation* defendants argued that, with more than 2 million documents to review, it would take reviewers more than 20,000 hours to perform the task – 10 man-years of billable time. But with predictive coding, it would take less than two weeks at a cost of roughly 1% of the cost of manual, human-review. In April 2012, the court ordered that defendants could proceed with the use of predictive coding for processing and production of electronically stored information (ESI).

The *Global Aerospace* decision stopped short of an unqualified approval of predictive coding. For example, predictive coding cannot work effectively if a representative corpus is not used for the initial training. The *Global Aerospace* court noted that the receiving party was free to challenge the completeness of the contents of the production and the manner in which predictive coding was used for new documents.

In the 2011 New York case *Da Silva Moore v. Publicis Groupe LLC, et al.*, the defendant proposed using predictive coding technology to cull the responsive documents from a collection comprising more than 3 million documents. The plaintiffs objected to the methodology employed by the defendants. The Hon. Andrew J. Peck of the Southern District of New York stated in his *Da Silva Moore v. Publicis Groupe* decision that predictive coding “is not a magic, Staples Easy-Button, solution appropriate for all cases.”

Peck delivered a useful definition for this technology, also known as computer-assisted coding. “By [predictive coding], I mean tools (different vendors use different names) that use sophisticated algorithms to enable the computer to determine relevance, based on interaction with (i.e., training by) a human reviewer.” Peck continued, “This judicial opinion now recognizes that computer-assisted review is an acceptable way to search for relevant ESI in appropriate cases.” This sanction of the use of predictive coding was a judicial first.

In Illinois, the plaintiffs in *Kleen Products, LLC, et al. v. Packaging Corporation of America, et al.* moved to force the use of predictive coding by the defendants to produce a more accurate production than the documents that were collected using the more traditional method of keyword searching. This motion was made moot in August 2012 by an agreement between the parties to forgo the use of predictive coding at that time.

Predictive coding has been used in many cases that have received no attention simply because there was no dispute about its use.

ment tool in e-discovery, holds promise for wider use in the records and information management arena.

Records professionals should reach out to e-discovery colleagues and technology partners to explore

the prospects for utilizing this technology. **END**

Doug Smith can be contacted at dsmith@wileyrein.com. See his bio on page 46.