FROM OCTOBER, *2010 LAW TECHNOLOGY NEWS:*

Crash or Soar?

Will the legal community accept "predictive coding?"

By Anne Kershaw & Joseph Howie

In the beginning there was brute force linear review of electronic data — lawyers would start at one end of a document collection and march through it, document by document, making decisions regarding relevance, confidentiality, privilege, topic, and importance.

Then came linear "clustered" review, where documents on similar topics are electronically identified and grouped so that review lawyers could read topically-related documents together and hopefully in sequence — clearly a huge step towards consistency.

Data volumes have continued to increase, causing further pressure to lower the cost of pre-production reviews. Litigants are finding relief from escalating costs with predictive coding, a human and technical process where a subset of records is examined by lawyers, and decisions made on those records are then propagated throughout the document population. This reduces or eliminates the need to examine all records.

Can litigants reliably produce documents that they haven't read? Our eDiscovery Institute recently conducted a survey and the results document that yes, you can — and save money in the process. We surveyed 11 e-discovery vendors who use predictive coding. The results report that, on average, predictive coding saved 45% of the costs of normal review — beyond the savings that could be obtained by duplicate consolidation and e-mail threading. Seven respondents reported that in individual cases the savings were 70% or more.

ADVANTAGES

The advantages of predictive coding extend beyond cost savings:

Transparency: All respondents track how their systems were used to select records. The types of data tracked varied with type of system used, but included items such as parameter settings, and relevance tags applied by experts. With linear review typically only the conclusion is recorded (e.g., relevant or not) with no insight into the decision-making process.

Replicability: Nine respondents said their system produces the same results on the same data if the steps outlined in the audit trail are followed. This is in marked contrast to human linear review, which typically produces low levels of agreement when different teams review the same records.

For example, in our study, "Document Categorization in Legal Electronic Discovery: Computer Classification versus Manual Review," Roitblat, Kershaw, Oot, *Journal of the American Society for Information Science and Technology* , 61(1):70–80, 2010), two teams reviewed 5,000 documents from a collection that had originally been reviewed for responding to a Department of Justice investigation. In terms of overall agreement (counting both responsive and nonresponsive documents), Team A agreed with the original reviewers 76% of the time, and with Team B 72% of the time. Team A agreed with Team B 70% of the time.

Considering only documents identified as responsive, Team A identified 48.8% of those identified by the original reviewers and Team B identified 53.9%. Of the documents identified as responsive by either Team A or B, the original reviewers identified 16.9%.

Two electronic service providers reviewed the complete collection using their document categorization systems, and achieved higher level of overall agreements with the original reviewers (83.2% and 83.6%) than those achieved by the human review teams. One identified 45.8% of the records originally identified as responsive, the other 52.7%

Reevaluating production sets: The costs of linear review are so high that parties rarely have the luxury of re-evaluating documents that have already been reviewed, regardless of what may have been learned about the issues after they were initially evaluated. By contrast, because predictive coding is based on human-assisted computer analysis, sets of documents can be examined multiple times using different parameters or sample sets.

Confidentiality: Because fewer reviewers see a given set of documents, predictive coding results in exposing confidential business documents to fewer people.

Shortened time lines: Computer-based analysis can review large volumes of data in short time frames, speeding discovery responses as well as internal and regulatory investigations. Faster delivery can be critical from a business perspective, e.g., obtaining regulatory approvals for a merger.

HOW IT WORKS

Survey respondents described how their processes worked. Six used queries as a component of predictive coding, and five used clustering — with some differences on whether terms could be inferred or not (e.g., whether a document that contained "Ford and Toyota" could find or associate documents that only contained the words "Chevy and Honda").

Some of the terms used to describe the respondents' offerings included "supervised learning" (Equivio), "linguistic statistical analysis (FTI), "machine learning" (InterLegis), "classification based technology" (Kroll), and "probabilistic latent semantic analysis" (Xerox).

OBSTACLES TO WIDER ADOPTION

Given the claimed advantages for predictive coding, why isn't everyone using it? The most mentioned reason, cited by 10 respondents, was uncertainty or fear about whether judges will accept predictive coding. (Paradoxically, at a recent U.S. Magistrates' Conference, a participant jurist asked for advice on how to convince lawyers to use this type of approach.)

The second and third reasons cited were lack of awareness of options on the part of in-house counsel, and insensitivity to costs of inefficiencies by law firms.

TERMINOLOGY

Eight respondents preferred some term other than "predictive coding" to describe the computerized approach to production review. Some took issue with "coding" as implying a level of precision that could be misleading. Some pointed out that their systems used a non-binary ranking system that required input to establish cutoff scores. Others suggested that "automated" or "propagated" would be more apt than "predictive."

E-MAIL THREADING

Respondents differed somewhat in how they used e-mail threading analysis in conjunction with predictive coding, i.e., whether predictive coding would treat all e-mails from a thread alike, or if some could be treated differently.

LANGUAGES

All of the respondents can process English, French, German and Spanish; eight can also handle Chinese, Japanese, Korean and Arabic.

TYPE MATTER/POPULATION SIZE

The respondents indicated that the value of predictive coding was higher in larger volume cases with short deadlines.

One respondent indicated it had highest value where the document population had been pre-culled, although another said that it was most beneficial where minimal initial document culling and classification had occurred.

Respondents varied in their responses to whether there was a minimum size below which predictive coding was less valuable. Two said 5,000 documents, another said 25,000 documents, while yet another reported a perception that it may be unnecessary in any case involving less than 25 GBs of electronically stored information.

QUALITY CONTROL

Eight respondents reported sampling nonselected records as a way of validating the results either routinely or as an option. The mere mention of sampling shows that predictive coding raises the bar on the level of technical competency required by the producing party — having enough knowledge to assess the results through sampling and other means.

To be comfortable that predictive coding technology satisfies the legal standards for document review (reasonableness), counsel undeniably will need to understand how the results are tested and verified. However, litigants should take comfort in knowing that if the same statistical rigor were applied to traditional linear review, it would often fail.

Study Participants

Capital Legal Solutions: **www.capitallegals.com**

Catalyst Repository Systems: **www.catalystsecure.com**

Equivio: **www.equivio.com**

FTI Technology: **www.ftitechnology.com**

Galivan Gallivan & O'Melia: **www.digitalwarroom.com**

Hot Neuron: **www.cluster-text.com**

InterLegis: **www.interlegis.com**

Kroll Ontrack: **www.krollontrack.com**

Recommind: **www.recommind.com**

Valora Technologies: **www.valoratech.com**

Xerox Litigation Services: **www.xerox-xls.com**

Anne Kershaw is principal of A. Kershaw Attorneys and Consultants and co-founder of the eDiscovery Institute, and is based in Tarrytown, N.Y. E-mail: Anne@AKershaw.com. Joseph Howie is EDI's director of metrics development and communications. E-mail: Joe@eDiscoveryInstitute.org. The report is available free at ***www.eDiscoveryInstitute.org****.*