Out Predict. Out Perform.

Using Built-In Sampling to Overcome Defensibility Concerns with Computer-Expedited Review

DESI IV Position Paper

Howard Sklar
Senior Counsel
Recommind, Inc.

For more information:

www.predictivecoding.com
www.recommind.com

Introduction

Linear document review – where individual reviewers manually review and "code" documents ordered by date, keyword, custodian or other simple fashion – has been the accepted standard within the legal industry for decades. However, time has proven this method to be notoriously inaccurate and very costly. And in a business environment where the sea of information – and therefore potentially relevant electronically stored information (ESI) – is ever-expanding, technology-enhanced methods for increasing the efficiency and accuracy of review are becoming an ever-more-important piece of the eDiscovery puzzle.

Courts have begun to push litigants to expedite the long-overdue paradigm shift from linear manual review to computer-expedited approaches, including Predictive Coding™. Judge Grimm framed this shift to computer-expedited review perfectly in a recent webinar[1]:

> *"I don't know how it can legitimately be said that manual review of certain data sets…can be accomplished in the world in which we live. There are certain data sets which I would say cannot be done in the time that we have as simply as a matter of arithmetic. So, the question then becomes what is the best methodology to do this. And this methodology is so much more preferable than keyword searching. I don't know what kind of an argument could be made by the person who would say keyword searching would suffice as opposed to this sophisticated analysis. That's just comparing two things that can't legitimately be compared. Because one is a bold guess as to what the significance of a particular word, while the other is a scientific analysis that is accompanied by a methodology…"*

The volume of ESI continues to grow at alarming rates and despite improved culling and early case assessment strategies[2], linear review remains too expensive, too time consuming and is, as articulated best by Judge Grimm, simply not feasible in many cases[3]. An AmLaw 50 law firm recently estimated that document review costs account for roughly one-half of a typical proceeding's budget[4]. However, new computer-expedited review techniques like Predictive Coding can slash that number[5] and provide a methodology that not only keeps budgets in check but speeds the review process in a reasonable and defensible manner.

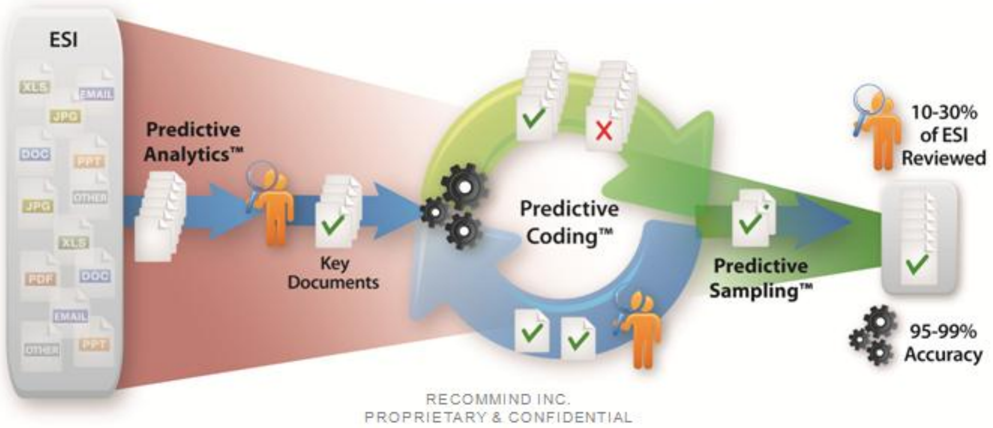
Predictive Coding addresses the core shortcomings of linear document review by automating the majority of the review process. Starting with a small number of documents identified by a knowledgeable person (typically a lawyer, but occasionally a paralegal) as a representative "seed set", Predictive Coding uses machine learning technology to identify and prioritize similar documents across an entire corpus – in the process literally "reviewing" all documents in a corpus, whether 10 megabytes or 10 terabytes. The result? A more thorough, more accurate, more defensible and far more cost-effective document review regardless of corpus size.

Unlike other computer-expedited offerings, however, Predictive Coding is not a "black box" technology where case teams are confronted with trying to explain the algorithms of an advanced search or application to a judge. Instead, Predictive Coding utilizes a workflow which includes built-in statistical sampling methodology that provides complete transparency and verifiability of review results that not only satisfies the Federal Rules' requirements for "reasonableness" of review process[6], but greatly exceeds linear review with respect to overall quality control and consistency of coding decisions.

The Process

The Predictive Coding starts with a person knowledgeable about the matter, typically a lawyer, developing an understanding of the corpus while identifying a small number of documents that are representative of the category(ies) to be reviewed and coded (i.e. relevance, responsiveness, privilege, issue-relation). This case manager uses sophisticated search and analytical tools, including keyword, Boolean and concept search, concept grouping and more than 40 other automatically populated filters collectively referred to as Predictive Analytics™, to identify probative documents for each category to be reviewed and coded. The case manager then drops each small seed set of documents into its relevant category and starts the "training" process, whereby the system uses each seed set to identify and prioritize all substantively similar documents over the complete corpus.[7] The case manager and review team (if any) then review and code all "computer suggested" documents to ensure their proper categorization and further calibrate the system. This iterative step is repeated until no further computer suggested documents are returned, meaning no additional substantively similar documents remain in the "unreviewed" portion of the corpus. The final step in the process employs Predictive Sampling™ methodology to ensure the accuracy and completeness of the Predictive Coding process (i.e. precision and recall) within an acceptable error rate, typically 95% or 99%. The result in most cases is a highly accurate and completely verifiable review process with as little as 10% of a corpus being reviewed and coded by human reviewers, generating dramatic cost and time savings.

Predictive Coding is based on the three (3) core workflow steps as follows:



1. **Predictive Analytics**: Predictive Analytics includes the use of keyword, Boolean and concept search, and data mining techniques – including over 40 automatically

populated filters – to help a case management team develop understanding of a matter and quickly identify sets (batches) of key documents for review. These sets are reviewed by the case team and establish seed documents to be trained upon during Predictive Coding's Adaptive ID Cycles (iterations).

2. **Adaptive ID Cycles**: Adaptive ID Cycles, also called iterations, are multiple occurrences of category training that identify additional documents that are "more like" seed documents. In this process, documents identified as being probative of a category during human review and Predictive Analytics are trained upon, with the application retrieving and prioritizing additional documents that it considers to be relevant to such category (i.e. substantively similar to the seed set). The cycle is as follows:

    a. Relevant seed documents are 'trained' upon
    b. The system suggests documents that are substantively similar to the seed set for such category
    c. Case team reviews/codes the suggested documents, providing further calibration for the system
    d. All relevant seed documents are 'trained' upon, and the iterations continue

3. **Predictive Sampling**: Predictive Sampling is the use of statistical sampling as a quality control process to test the results of a Predictive Coding review. It provides quantifiable validation that the process used was reasonable and, as a result, defensible. Predictive Sampling is used after Adaptive ID Cycles yield no or a very small amount of responsive documents, meaning no substantively similar documents remain unreviewed and uncoded. The process entails pulling a random sample of documents that have not been reviewed and placing them under human evaluation for responsiveness. The review can be deemed complete after quality control sampling is verified to provide a statistical certainty in the completeness of the review.

Predictive Sampling Examined

Quality control in the document review process has long been identified as something which is at best unevenly applied and at worst nonexistent.[8] Of particular concern – and criticism by no less than the Sedona Conference[9] – has been the reliance on such inaccurate tools as keyword search. As such, Landmark eDiscovery cases including the Victor Stanley[10] and Mt. Hawley Insurance Co.[11] decisions have pushed parties to not just embrace more advanced technology, but have gone so far as to identify sampling as the only prudent way to test the reliability of search, document review and productions irrespective of technology or approach utilized.

In keeping with this emerging judicial mandate, the Predictive Coding workflow automates the sampling process in the form of Predictive Sampling, which provides statistically sound certainty rates for responsiveness, issue relation, etc. The soundness of this approach has been corroborated by eDiscovery industry commentators, including Brian Babineau, Vice President of Research and Analyst Services with Enterprise Strategy Group,

> *"Predictive Sampling assesses the thoroughness and quality of automated document review, helping to fortify the defensibility of Predictive Coding. Leading jurists have already written that the superiority of human, eyes-on review is a myth, so law firms continue to work with technology vendors to fill in much of this gap. Predictive Coding with Predictive Sampling enables users to comfortably leverage technology to attain a level of speed and accuracy that is not achievable with traditional linear review processes."*

The Predictive Sampling process is relatively straightforward. A statistically significant number of documents (typically 2,000 – 10,000 for statistical significance) are randomly set aside by the system before the review or analysis process begins; this set of documents is the "control set" against which the review – both by the review team and the Predictive Coding system – will be measured to validate the accuracy and error rate of all coding decisions. This control set is reviewed by the case team for all relevant categories, i.e. relevance, responsiveness, privilege and/or issue relation, with the positive/negative rates for all such categories automatically tracked by the system.

Once the Adaptive ID Cycle step is completed, a small selection of the remaining, unreviewed corpus is randomly selected by the system for review by the review team (again, typically 2,000 – 10,000 documents for statistical significance). This latter set is then reviewed and coded to see if any probative-yet-unidentified documents (aka false negatives) can be found. The results of this review are then compared against the results from the review of the initial control set, from which a statistically significant and verifiable measurement of the Predictive Coding process's accuracy and completeness (i.e. precision and recall) are verified.

Incidentally, while beyond the scope of this paper it has been shown that the above process has a rather significant benefit beyond the validation of the Predictive Coding process: the ability to use quality control in the review process as an offensive weapon.

Unparalleled Review Speed, Accuracy, Cost Savings *and* Defensibility

The most immediate benefits of Predictive Coding are the dramatic reduction in review time required, thereby decreasing review costs significantly while simultaneously improving review quality. Predictive Coding has been shown to speed up the review process by a factor of 2-5x, yielding 50-90% savings in the cost of review. Time and cost improvements include:

- Predictive Analytics provide early insight into the substance of a corpus and key documents before review has begun. This allows a targeted approach to creating seed documents to be used for category training.
- More relevant documents are in front of reviewers, more often and more quickly, leading to reviewers seeing less non-relevant documents thereby further expediting the review process.
- The process provides a pre-populated (predictive) coding form to the reviewer. The human review is mostly a confirmation of computer-suggested coding, which thus saves review time and improves coding consistency.
- The process provides highlighting hints within the document to guide the reviewer in his/her decisions, and thus to quickly focus his/her attention on the most important parts of the document – which is particularly helpful with longer documents.
- Category training provides a self-assessment of quality in terms of a confidence score. This allows the reviewer to focus on the most critical parts of the review.

Additional improvements in review quality with Predictive Coding enhance and improve coding decisions made by case teams:

- The predictive suggestion in the coding form leads to a significantly more consistent review across different reviewers.
- The human reviewer is typically very precise whenever making a positive decision. However, the completeness of the reviewer's coding is typically lacking. For example, reviewers may miss certain issue codes, not becoming aware of sections in a document that lead to privilege classification, etc. Predictive Coding will not only provide a predictive check for reviewers to investigate but also provides highlights to critical concepts identified on the document. Thus alerting reviewers to critical aspects of documents.
- Typically, category training is run in a mode that is overly complete, i.e. errors on the side of recall. As a result, the overall review quality typically improves significantly, while maintaining a 2-5x speed improvement.
- Predictive Sampling used as a quality control process can provide case teams with a 95-99% certainty that relevant documents have been identified, confidence that is unmatched by any linear review or keyword search method.

Conclusion

In an era where escalating costs and increasing volume dictate a better way to manage the document review process, more and more legal teams are turning toward new methodologies to address client needs and concerns. The question is no longer if legal teams must reduce the time and cost of review but what method will they implement that is effective but also defensible. In response to this acute need, Predictive Coding with Predictive Sampling has achieved the "holy grail" of document review: the judgment and intelligence of human decision-making, the speed and cost effectiveness of computer–

assisted review, and the reasonableness and defensibility of statistical sampling. This patented methodology facilitates a fully defensible review while dramatically reducing review costs and timelines, as well as improving the accuracy and consistency of document review.

[1] Webinar found at http://www.esibytes.com/?p=1572. Cited reference at 40:10. Last accessed on April 15, 2011.
[2] *Jason Robman*: The power of automated early case assessment in response to litigation and regulatory inquiries. The Metropolitan Corporate Counsel, p33, March 2009.
[3] *Craig Carpenter*: Document review 2.0: Leverage technology for faster and more accurate review. The Metropolitan Corporate Counsel, February 2008.
[4] Anonymous AmLaw 100 Recommind customer, January, 2011.
[5] Robert W. Trenchard and Steven Berrent: Hope for Reversing Commoditization of Document Review? New York Law Journal, http://www.nylj.com, p3, April 18, 2011.
[6] Robert W. Trenchard and Steven Berrent: The Defensibility of Non-Human Document Review. Digital Discovery & e-Evidence, 11 DDEE 03, 02/03/2011.
[7] *Craig Carpenter*: E-Discovery: Use Predictive Tagging to Reduce Cost and Error. The Metropolitan Corporate Counsel, April 2009
[8] *Craig Carpenter*: Predictive Coding Explained. INFOcus blog post, March 10, 2010
[9] See Practice Point 1 from The Sedona Conference Best Practices Commentary on the use of Search and Information Retrieval Methods in E-Discovery.
[10] Victor Stanley Inc. v. Creative Pipe Inc., --F Supp 2d--, 2008 WL 221841, *3 (D. Md. May 29, 2008).
[11] *Mt. Hawley Ins. Co. v. Felman Prod. Inc.,* 2010 WL 1990555 (S.D.W.Va. May 18, 2010).