

White Paper

Predictive Coding

The Next Phase of Electronic Discovery Process Automation

By Katey Wood and Brian Babineau

August, 2011

This ESG White Paper was commissioned by Recommind and is distributed under license from ESG.

Contents

Introduction	3
Automation in the Legal Sector	3
Moving Beyond Paper	3
Beyond Linear Review	4
The Promise of Search and Analytics for Attorney Review	4
Predictive Coding – the Next Step	5
Do Judges Believe in Magic?.....	5
Know Your Tools	5
Automated Approaches – the Nuts and Bolts	6
How to Use It, How to Choose It	7
The Bigger Truth	8

All trademark names are property of their respective companies. Information contained in this publication has been obtained by sources The Enterprise Strategy Group (ESG) considers to be reliable but is not warranted by ESG. This publication may contain opinions of ESG, which are subject to change from time to time. This publication is copyrighted by The Enterprise Strategy Group, Inc. Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of the Enterprise Strategy Group, Inc., is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact ESG Client Relations at (508) 482-0188.

Introduction

Fulbright & Jaworski's 7th annual litigation trends report cited, for the first time in the history of the survey and by a large margin, electronic discovery as the top area for increased litigation spending over the next year. It's clear that corporate litigants continue to grapple with the cost implications of the 2006 changes to the Federal Rules of Civil Procedure, which officially made all electronically-stored information discoverable in US courts. According to a 2010 Duke University survey of litigation costs at major companies, in 2006-2008, the average discovery costs per case ranged from \$621,880 to \$2,993,567, topping \$2,354,868 to \$9,759,900 per case for companies at the high end of the spectrum.¹

The continued growth of e-discovery expenses in 2011 is even more startling in light of the enormous amount of technological innovation and solution development over the last five years specifically designed to *reduce* associated expenses. The number of matters involving electronic discovery, coupled with the amount of data that could be discoverable during any of these cases, is obviously growing at rates much faster than organizations and their corporate counsel can manage cost-effectively. Many corporate litigants have responded by bringing more of the process in-house and better automating the more labor-intensive elements, particularly in the identification, collection, preservation, and processing of electronically stored information (ESI) at the earliest stages of litigation response.

Yet these in-house efforts alone have not adequately addressed the unsustainable amount of data still slated for downstream attorney review, much less the cost inefficiencies and redundant efforts involved. Duke's litigation cost survey estimated that in major cases going to trial in 2008, the ratio of pages discovered to pages entered as exhibits at trial was as high as 1000/1. On average, 4,980,441 pages of documents were produced, but only 4,772 exhibit pages were marked. Traditional linear attorney review, typically the most costly and effort-intensive phase of discovery, is not practical or sustainable given these high data volumes and low proportions of relevant information.

While many phases of litigation and discovery have been optimized for efficiency on a large scale through automation, attorney review remains a "final frontier" of more advanced technical methods. Automated information retrieval has been used for decades to navigate the large-scale, complex data volumes of the computer age. In fact, several e-discovery vendors and service providers have developed their own techniques and workflows to incorporate more advanced search and navigation into attorney review. These, however, have faced an adoption lag due to technical inertia in the legal sector, overall lack of judicial guidance on coping proactively with the problems of discovery, and arguments of methods and defensibility over human approaches despite studies showing that human review itself only achieves about 50% consistency.²

One such e-discovery vendor, [Recommind](#), recently patented its Predictive Coding methods for iterative computer-assisted document analysis and review, potentially creating more critical mass in the market around efforts to accelerate and optimize document review. This paper will address the necessity of predictive coding in the context of broader trends in legal sector automation and outline considerations for comparing and evaluating different approaches in the market.

Automation in the Legal Sector

Moving Beyond Paper

The history and practice of law are fundamentally human endeavors, refined and studied through much of recorded history. Today, the legal sector remains one of our most heavily expertise-driven industries, depending on human intellect to analyze and negotiate complex, nuanced issues with both theoretical and practical lenses.

Technology and automation play essential roles in enabling the practice of law just as they do in any modern business. The legal libraries which once lined law firm office walls are now more readily accessed and searched

¹ Source: Duke University, *Litigation Cost Survey of Major Companies*, 2010.

² Anne Kershaw, "Automated Document Review Proves Its Reliability," *Digital Discovery and e-Evidence*, 2005.

through Internet research portals like WestLaw and LexisNexis. Evidence for attorney review, once stacked in heavy boxes of paper files, is now scanned or processed and loaded into review software for more organized, efficient, and portable review on computers. Depositions can now be streamed, searched, and marked up in real time thanks to more automated transcript management. Time capture applications help track and record billable hours, even from mobile devices, to prevent billing “leaks” and enable attorneys to focus on more important tasks.

The legal industry is evolving to new digital operations that don’t pre-date computers, an inevitable step many businesses are taking in order to survive the 21st century and even gain advantage from the modern “digital deluge” of data and information. None of these innovations has resulted in the end of lawyers or their craft; instead, they allow law firms and attorneys to do their jobs better, more efficiently, and with less drudgerous, time-consuming grunt work.

Beyond Linear Review

This switch to more digital methods has already begun in order to process large amounts of data within achievable time frames. While paralegals once did manual data entry and coding in order to enter paper evidence into a database, the introduction of optical character recognition (OCR) and auto-coding technology lets computers capture document text automatically from paper, extract entities from it, and auto-populate bibliographic metadata codes e into a database for easier searching and organization. Taking the time for humans to perform this task on all evidence would be unthinkable in many cases. The same is now true of traditional linear attorney review in which attorneys methodically read and code every document in evidence, either in chronological order or in no order at all.

Fundamentally, computers do not replace humans or “think” like them, but there are specific tasks they can be programmed to perform more quickly and efficiently. And the benefits of computers and automation extend far beyond the efficiency of digital methods compared to paper, particularly in search.

The Promise of Search and Analytics for Attorney Review

Digitization, computing power, and the Internet have allowed us to store, process, and access information at much higher volumes and distances, but it is *search and associated technologies* that have truly unlocked the promise of finding what we need to make it useful. Internet search engines are a good example: consumers now turn to a single search box for information they used to get from a phone book, newspaper, travel agent, shopping trip, or library research. They have come to depend on search algorithms for retrieving this information, saving time and effort by automatically “crawling” vast amounts of data to find what they want.

Time and human effort are especially valuable in the practice of law—in fact, with the billable hour, time literally is money to law firms. But attorney review is a very different scenario from conventional information-seekers looking for specific information or news on the web. Attorney reviewers are typically given a document set with few ideas about what might be in it and must find not only the “best match” for relevance, but *everything* of relevance. Attorneys must be able to quickly navigate a large document collection, determine its contents, and locate what is needed for the case at hand with both high precision and recall.

There is no silver bullet for this task. Instead, a number of different innovations in search and analytics have been introduced in the legal context over the years to streamline the process.

1. In attorney review tools, search is commonly available on keywords, letting users find specific words, phrases, or even strings or patterns of characters and “wild cards.” Metadata is also used to search and sort the collection based on the author, format, date of creation, and other facets of the data object.
2. Innovation in data analytics has given computers even more tools to organize information usefully. Deduplication and similar technologies are used for identifying, comparing, and potentially culling similar documents. E-mail threading and network analysis can reconstitute the structure of conversations and create maps of social networks between correspondents.

3. Concept search can be used to automatically group documents into similar conceptual areas for greater accuracy over keyword-only search by quickly finding documents conceptually similar to each other, regardless of the keywords used.

Used effectively, these methods can save thousands of man-hours of work, improving attorneys' performance in quickly and accurately narrowing down a collection to find relevant documents. They represent an evolution in how computers are able to serve the legal use case by processing text, organizing and searching a collection of data, and even imposing more conceptual structure on it. They also represent great refinement in the tools available to attorneys to perform their tasks.

Predictive Coding – the Next Step

Do Judges Believe in Magic?

The next logical step in this evolution is harnessing computing power to predict attorney coding decisions. Looked at as a progression—from keyword search, to culling duplicates, to grouping related documents, to predictive coding—it is tempting to say that we're reaching a point where computers are better able to "learn" or "understand" the attorneys' tasks, particularly when the results given by vendors for cutting data volumes and costs make the process seem automatic and almost "magical." But that isn't accurate. Machines are not able to "understand" or "think" like humans; they can only be programmed to perform an assigned task under specific conditions within a certain level of accuracy.

In terms of defensibility, judges have given little formal guidance, but have made clear they have a general expectation of reasonable and defensible methods—not perfection. Parties should be prepared to explain their methods and why they chose them as appropriate for the task, show that the process was properly implemented, and prove that it was subjected to testing for quality assurance. For this reason, it's important to understand how these technologies work and are being applied in a given tool. There is no magic in these systems, nor are they virtual "robotlawyers." They can be powerful tools for attorneys, but only if incorporated into the review process with some understanding of how they work as well as their strengths and weaknesses.

Know Your Tools

"Predictive coding" is a function, not a specific technology; so the technical methods, process, and workflow behind different vendors' underlying search and text mining may vary. Because methods used in e-discovery must be defensible in court, it's vital that attorneys have some understanding of how their tools function and what kind of results to expect from them. How does the tool find things and how well does it work? Because there are few technical standards in e-discovery, some approaches are more sophisticated (and defensible) in various scenarios than others. The burden is on the user to understand the difference between them and determine which might be useful in a given situation.

For example, vendors may use a "rules-based" method in which documents are retrieved based on pre-determined criteria gleaned from a knowledgebase of queries effective in similar investigations. This could be as simple as a keyword search for "confidential" to locate potentially privileged documents or as complicated as Boolean queries dozens of terms long. In other instances, a simple "concept" search used to find and cluster similar documents containing different words with similar meaning could be based on nothing more than a Wordnet thesaurus of English-language synonyms—like car, automobile, vehicle, etc.

These approaches can be useful in searching, filtering, and culling a document collection, particularly when its contents are known and the object of the investigation is understood (i.e., investigators know what they are looking for). Multiple iterations and human refinement can improve results when operated by an experienced investigator. However the challenges of legal document search—large, diverse collections of unknown documents in multiple formats, sometimes in several languages, and with few hints for finding less intuitive code words or project names—have increasingly lent themselves to new approaches. While linguistic and keyword methods can be effective in retrieving specific information from "known" document sets when used by experienced users, they can

break down under the burden of “big data” and language ambiguity, often producing false positives while missing relevant information—particularly under the realities of litigation timeframes. Keyword search alone limits the power of tools by relying on the user to make informed queries and the index to provide exactly what the user is searching for. Alternate methods of search, navigation, and aggregating “like” documents have emerged and been adapted to the legal use case over the last few decades based on non-linguistic context clues of word count, proximity, frequency, and other document characteristics.

Automated Approaches – the Nuts and Bolts

Computer algorithms for information retrieval have been in use for decades in the legal field, and have improved for the legal use case over the years both in technical accuracy and process. Search algorithms can vary in how they handle the challenges of legal data sets, namely: their size, sometimes unknown contents, the heterogeneous nature of the data, and the “dirtiness” of data from misspelling, scanned OCRs, incomplete documents, or other “noise” that subverts precise retrieval. While there are many tweaks and versions of algorithms for information retrieval, here we’ll outline some of the more popular ones used in the legal space.

Some automated coding tools categorize documents through mathematical algorithms such as latent semantic indexing (LSI). LSI can use a set of example documents to categorize a larger body of documents. It does this by using concepts in the example set to identify terms and concepts used similarly in context throughout the larger corpus and categorize the documents accordingly. Or LSI can “cluster” documents without a training set, according to similarities between them in how concepts appear in context. Typically, the LSI approach will realize improved precision over Boolean search, but does not completely address recall.

Automated statistical methods like probabilistic latent semantic analysis (PLSA) are based on statistical analysis of word context and *occurrence* to find patterns and identify concepts—the difference from this and LSI being the probability of variables co-occurring (for example, the chance of given documents containing the same words, or using them in the same context). The PLSA approach is able to group documents based on content containing similar concepts and can do so even in the absence of taxonomies and other category information. Although training documents can be used in the process, they are not required to satisfactorily address a combination of both precision and recall.

In the last few years, some review tools have incorporated prioritization technology and workflow in order to locate relevant documents within a large collection and have attorneys start reviewing them first. These tools are based on machine learning and rely on iterative “training” sets coded by humans to “seed” the system. Much as one “trains” a spam filter to better identify (and block) spam by pointing out its mistakes when they arrive in the inbox, these applications iteratively learn from human judgment to retrieve and prioritize documents that fit a certain coding criteria to “find similar” results once an attorney has determined a training document to be responsive or privileged.

These tools need a sufficient sample set of documents coded by a human in order to find items within a certain degree of similarity; that is, they must be “trained” to categorize documents with a statistical probability of accuracy within a desired confidence range. The size of these training sets varies by a number of factors depending on the tool, the desired confidence level, and how the technology is being used.

Statistical approaches have the benefit of being language-independent. They are typically able to find related concepts in documents whether or not they contain the same terms. In fact, they can work with many different types and structures of documents, and also frequently work on data in multiple languages.

The accuracy of statistical approaches can be tested through statistical sampling. Basically, a small set of documents is randomly selected from the universe of all documents for which we want to test a hypothesis or determine an error rate. The small set is then batched out for manual review to determine an empirical error rate. The margin of error around the empirical estimate can then be directly related to the sample size and is, most importantly, independent of the actual training technique and can hence be used as a defensibility methodology for a broad range of automated methods.

Of course, the same sampling techniques can be used to validate review results from human reviewers if a linear process were followed. It is well known that human review is far from perfect and error rates from 10-50% (or more) are common in large case linear reviews. In addition, training techniques can be very effectively used to detect potential human coding errors by using the coding result of linear review as a starting point for “training.”

How to Use It, How to Choose It

Once trained, predictive coding tools can suggest coding categorizations for the remaining documents in a collection. Documents with a higher probability of relevance can be prioritized for human review while the rest can be sampled to verify that they are unlikely to contain relevant documents and don't necessarily need to be reviewed by humans. Predictive coding can be used not just for the review portion of a case, but also early case assessment, regulatory review, and other scenarios. If documents have already been reviewed on a case by humans, predictive coding tools can also sometimes be used to QC them and find possible errors.

Understanding the potential use cases for these tools is vital, including as much detail as possible on how it will be used in practice.

- What are the usage and setup terms and costs of the system?
- Can it be operated by existing GC or law firm staff?
- Is it compatible with existing systems?
- How long does it take to set up the system?
- How does it categorize documents, and is its approach linguistic, statistical, or a hybrid?
- What is the desired confidence level in the accuracy of results for a given matter, and how does that effect timeframes for completion?
- What are examples or particular case studies of how many pre-coded documents were needed and timeframes necessary to train the system to an acceptable confidence level in various investigations?
- What kind of “seed” documents are required for the training set (responsive, unresponsive, mixed)?
- Can the tool be adjusted or refined for new or iterative document sets being added to the collection?
- At what volume does this approach become most cost- and time-efficient?
- What are the tool's documentation and reporting capabilities for defending its methods in court?
- Does it include integrated statistical sampling as a QC measure to verify its accuracy?

Product options and pricing vary substantially, making it vital to establish organizational requirements early and engage in a proof of concept with potential vendors.

The Bigger Truth

The preceding passages outline how technology has promoted change and evolution in how we practice law, making it more adaptive to current circumstances and the problems faced by clients. This change has been a slow progression over the years, but recent events have made the need more acute. Businesses already struggling with a growing flood of electronic data were dealt another blow with the 2006 changes to the Federal Rules of Civil Procedure, which officially made all electronic data discoverable in US civil proceedings, vastly enlarging the scope of potential litigation requests. Meanwhile, the aftermath of the recession has cultivated more litigation and regulatory activity in many cases, but also created pricing pressure in responding to it.

In helping clients navigate the new challenges of a changing legal environment, law firms can either become a trusted ally offering competitive differentiation from their peers or they can end up as one more scapegoat in the crisis their clients are facing. Either way, they cannot continue conducting business as usual.

It is no longer possible for human reviewers to lay eyes on all of the terabytes of data sometimes produced for review. Outsourcing to contract reviewers can make it more affordable in some instances, but this is only a band-aid in addressing the ever-growing volumes of data. Just as we wouldn't hire a million librarians to vet every link on the web to find the most relevant one to our needs or pay a million paralegals to enter and code documents into a database, we cannot hire attorneys to review every document in a terabytes-large collection. What's more, humans are not infallible and can be even less consistent than machines at performing repetitive document review tasks for long hours under tight deadlines.

Clients demanding greater accountability from their law firms cannot tolerate the financial burden. Moreover, as stronger project management and QC controls take hold in the legal field with new tracking, metrics, and measurement assigned to monitor reviewers' work, clients are increasingly demanding better results and performance which will increasingly only be achievable through more automated methods such as predictive coding.



Enterprise Strategy Group | **Getting to the bigger truth.**