

**ABA SECTION OF LITIGATION 2012 SECTION ANNUAL CONFERENCE
APRIL 18-20, 2012:
PREDICTIVE CODING**

Predictive Coding

SUBMITTED IN SUPPORT OF THE PANEL DISCUSSION

INTRODUCTION

Technology has created a problem. The ease of use, the ubiquitous availability, the ease of retention versus the effort of management has caused emails (and other communications) to build into an overwhelming volume of data that is exponentially more expensive to deal with in litigation. Can technology solve the problem that technology created? Promoters of predictive coding claim that it can. That's a big promise, because parties involved in litigation tell us that this is a big problem.

In order to understand the promise, it's important to understand the problem, to build a common vocabulary, to understand the technology, and to see how it's been applied within a legal context. This paper strives to do just that, which is difficult because, as we'll see, this is an area that is 50+ years young and there is still a lot left to do...

THE PROBLEM

On Sept. 9, 2011, the Discovery Subcommittee of the Advisory Committee on Civil Rules held a mini-conference on preservation and sanctions in Dallas, Texas.¹ During that mini-conference there were submissions and testimony from a variety of corporations, consultants, lawyers, judges, and academia. One of the facts presented by Microsoft was the statistic that for every 340,000 pages of information preserved for litigation; only 1 is actually used in litigation.² While Microsoft may not be representative of every company, or have information representative of every litigation, other comments during the mini-conference echoed similar experiences.

¹ Notes from the Mini-Conference on Preservation and Sanctions, September 9, 2011.

² Microsoft letter to the Chair of the Advisory Committee on Civil Rules, August 31, 2011, p. 4.

How does litigation deal with such a huge volume of source information (340,000) in a cost effective manner that allows the final information (1) to surface?

As for the costs beyond the cited experience of Microsoft, it has been estimated that 50% of the cost of litigation is discovery, and that 50% of the cost of discovery is related to electronically stored information (ESI).

And the problem is not unique to the practice of law. The organization and retrieval of information has been an elusive problem ever since people started creating and storing information. ARMA International (originally known as “The Association of Records Managers and Administrators”) was founded in 1955 and is indicative of the early need for managing records – which includes retrieving information – during the computer age.³ More recently, as early as 1992 the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense (DoD) began a co-sponsored conference referred to as TREC (the “**T**ext **R**etrieval **C**onference”) which has been supporting research within the information retrieval community by providing the infrastructure necessary for the evaluation of text retrieval methodologies.⁴ TREC operates several different tracks which provide an insight into the varied nature of the problems faced. These tracks include:

- Contextual Suggestion Track
- Crowdsourcing Track
- Knowledge Base Acceleration Track
- Legal Track
- Medial Records Track
- Microblog Track
- Session Track
- Web Track

There are another 17 past tracks, as well. Clearly, even for a focused exercise, this is a big challenge.

IS PREDICTIVE CODING A SOLUTION?

There may be many potential solutions to the challenges of U.S. litigation: Better advocacy, presumptive limits, and potential rule changes are all in various stages of evolution. But these may take months, or years to implement.

Today’s technology solution is using computer technology to analyze large data sets for thematic patterns, learn the choices made by a highly focused litigation team, then apply those choices across the remaining corpus. There are multiple competing models to do this. Collectively, and colloquially, they are called “predictive coding”.

DEFINITIONS

Before we get into the main substance of predictive coding, it would be instructive to address some of the terminology up front. Here is a list of a few common terms. The list is grouped by

³ <http://www.arma.org/about/overview/index.cfm>

⁴ <http://trec.nist.gov/overview.html>

relationship, rather than alphabetically. This is because most of these terms are either linked to another term, or are alternatives or opposites of another term:

- Precision and Recall go together
- Linear Review and Non-Linear Review are alternatives
- Supervised Learning and Unsupervised Learning are alternatives
- Prioritized Review and Automated Review are alternatives

PRECISION

When a document selection is made, manually or by any other means, precision is the measurement within the selected set of the ratio of desired documents (relevant) to the total number of selected documents. This is a common measurement in trying to determine the effectiveness of Information Retrieval, or the relative effectiveness between two different systems applied to the same corpus. A higher precision is better. A lower precision indicates a larger number of false positives.

RECALL

When a document selection is made, recall is the measurement of the number of desired documents (relevant) within the selected set versus the total number of relevant documents within the entire corpus. This is another common measurement in trying to determine the effectiveness of Information Retrieval, or the relative effectiveness between two different systems applied to the same corpus. A higher recall is better. A lower recall indicates a larger number of false negatives.

LINEAR REVIEW

Linear review was the traditional methodology for document review. Arising out of paper document review, linear review does not leverage any knowledge about documents to improve the order in which they are reviewed. Through linear review, documents are reviewed in the exact order in which they were stored and delivered to the review team.

NON LINEAR REVIEW

Non-linear review is a review methodology in which efforts are made to leverage any information about a document or contained within a document to attempt to improve the review process. This may include grouping, sorting, sampling or other advanced technical approaches.

SUPERVISED LEARNING

Supervised learning is an algorithm that learns from human decisions and then has the ability to apply those decisions to new data. It is considered supervised because it uses a training set. The selection and development of the training set, the settings of the supervised learning algorithm, and the manner in which it is actually relied upon, are important considerations in supervised learning.

UNSUPERVISED LEARNING

Some algorithms can analyze data without the benefit of a training set. These algorithms seek to discern patterns already inherent within that data. Often, clustering is an example of unsupervised learning. In clustering, word occurrences are analyzed within documents to derive common themes. Documents can then be given weights reflecting how strongly or weakly they correlate to

a derived theme. It is considered unsupervised because the system derives the themes without specific human intervention.

An example is the derivation of “topics” or “clusters” from within litigation document collections. Those topics can then be used to group documents, find related documents, review documents, or even bulk tag documents during review.

PRIORITIZED REVIEW

Prioritized review is the use of a selection methodology to push suspected relevant documents to the front of the queue. An advanced use of prioritized review is to then use the actual human review decisions of those suspected relevant documents to revise the rankings of the remaining documents in an iterative (or continuous) process.

If the original ranking uses unsupervised learning and the subsequent ranking learns from the human decisions, this is sometimes referred to as semi-supervised learning.

AUTOMATED REVIEW

Automated review, as opposed to prioritized review, is where a technology is relied upon to make document selection designations without subsequent human confirmation. The algorithmic decision is accepted as final.⁵ Automated Review followed by an exhaustive human confirmation is more similar to a special, one-iteration case of Prioritized Review than it is of truly automated review.

SEARCH TERMS

Search terms (typically referred to as Boolean Search Terms but notably expanded beyond the original AND, OR and NOT operators) are textual strings that are used to search within a body of information. This can be done against documents that have been indexed, or against non-indexed documents. Both methods have their own pros and cons. Indexing takes time (and expense) on the front-end that can translate to faster results. This is typically a decision based on how often you expect to search the information. For example, if you are only searching a few times, or one time, such as when filtering information, it may be faster and more economical to skip indexing.

SAMPLING

Sampling of information can be very helpful because reviewing a portion of the information can be a cost effective way of making a decision against a larger population. Sampling has many different flavors that go beyond the scope of this paper. But some common categories of sampling include judgmental sampling (looking where you expect to find answers), hierarchical sampling (choosing samples at the group, collection, container or lot level), random sampling, and statistical and non-statistical sampling.

In the most recent discussions of predictive coding (and other forms of technology assisted review) sampling has been cited for improving the understanding of a document collection, for

⁵ Caution: the word ‘final’ can be misleading. Using a given technology to designate decisions within a document population, such as using email addresses of particular employees to select documents for future consideration, may result in a set of documents that are not subsequently reviewed again *for that decision* but may be separately subject to one or more review cycles (e.g. for responsiveness and or for privilege).

developing and testing search terms, for training supervised learning systems, even for the selection of representative custodians.

One very significant application of sampling has been in the validation of a given process. To validate processes, usually random samples are chosen that are substantive enough to make statistically valid assertions about the results. This can result in a statement such as – after the application of search terms to the document population to help locate potentially relevant documents, we tested the remaining documents by manually reviewing a specific number of documents (e.g. 384) which allowed us to give an estimate of the number of remaining potentially relevant documents (e.g. 14.3%), and to calculate the confidence level of that estimate (e.g. 95%), and to calculate an error range (e.g. +/- 5%) of that estimate.

SEARCH ENGINE OPTIMIZATION

Search engine optimization (SEO) is the science behind understanding the web page ranking systems of Google, Yahoo! and other search engines, and tweaking company web pages and embedded terms to maximize performance against specific searches.

Ultimately this is related to the information retrieval techniques that these search engines use. This is relevant to legal information retrieval because, as large as the legal problem is, the broader Internet challenge of information retrieval is a magnitude larger. For example, since Internet advertising and web traffic is an industry worth billions of dollars⁶, there has been a significant amount of investment into search engine development and search engine optimization development in support of advertising and web traffic attraction.

These investments in broad computer science applications are developing the same technologies that are being applied in the legal market.

HYPOTHETICAL CASE STUDY

The review of vocabulary is most useful when it is seen in operation. Therefore, it may be helpful to understand some of these technologies within a legal hypothetical.

ACME SUES BRASSO

Acme Ltd has been manufacturing The Acme Rocket Chair for 40 years. Brasso, Inc. has been purchasing and reselling Acme rocket chairs for most of that time under an agreement that it would only sell them within specific Western States within certain pricing guidelines. Acme has begun to suspect that Brasso is violating the marketing territory that they agreed to by selling Acme chairs in Eastern States and doing so at a discount price, and that those violations are affecting Acme's other resellers and Acme's ability to command a premium price for a premium product. Correspondence and diplomacy have failed, and Acme has sued Brasso for violating their resell agreement.

Brasso, which sells an entire portfolio of products nationwide, has 250 sales people on the East Coast that it considers as the most likely people to have potentially relevant documents. In response to a request for production, Brasso now needs to collect, process, review and produce documents to Acme.

⁶ As of March 2012, Google has a market cap of \$208 Billion USD.

Brasso uses judgmental sampling in determining that the most likely source of responsive documents will be the Chief Sales Officer (CSO), The Chief Marketing Officer (CMO) and members of the 250 person salesforce.

Brasso then uses hierarchical sampling within the salesforce to randomly chose 25 salespeople from which to collect documents (in addition to the CSO and CMO).

Brasso ends up collecting 60 GB of email. Brasso then develops a set of search terms to apply to that 60 GB of email.⁷ Those search terms result in 6 GB of hits. 6 GB can translate to 75,000 documents or more.

Brasso then decides to apply supervised learning to this data set. In doing so, they first review a random set of 5,000 documents from the 75,000, marking responsive documents. The results of that test are use to train a supervised learning engine.

The engine is then used to rank the remaining 70,000 documents. Brasso then reviews the next 5,000 documents that the system believes will be most similar to the documents just designated as responsive.

The engine is trained again, this time on the results of both sets. Brasso then reviews a third set of 5,000 documents. In theory, Brasso begins to see that the third set begins to have a lower percentage of relevant documents. Brasso repeats training, then reviews a fourth set only to find that the percentage of relevant documents has dropped very low. So low, that they feel that they may be finished identifying potentially relevant documents. If true, then they have reviewed 20,000 documents instead of 75,000 documents to prepare their production. If false, then they are not finished and there may still be a significant number of responsive documents remaining.

Brasso decides to test the remaining documents to estimate the number of remaining responsive documents. Brasso has 500 documents randomly chosen and reviewed. They find that 75 of those are responsive. Brasso calculates that they are 95% confident that there remains a residual 15% +/- 5% responsive documents.

Based on this information, the nature of the documents that they are finding, the facts of the case, and the anticipated cost, burden and delay of additional review, Brasso decides to stop reviewing documents. They conduct a privilege review of those documents identified for production. They make their production of the non-privileged documents.

Brasso now feels that they have been able to conduct a successful review by using technology to target 20,000 documents (plus testing of 500) rather than reviewing 75,000 documents. Depending on the cost of the system that they used, this may result in a significant cost savings to Brasso.

ASSESSMENT OF THE RESULTS

Did Brasso save money while meeting their obligations? Possibly. The hypothetical is an oversimplification of the due diligence that may go into a typical case. Also, Supervised Learning is still relatively new to the legal field and the prices (and the efficiency) vary quite a bit. Thus

⁷ There are many considerations when developing search terms within a legal context. That discussion goes beyond the scope of this paper.

any money that they save on the cost of human review has to be balanced with the additional cost of using technology. For this reason, litigants, such as this hypothetical, use a combination of technologies and techniques at different stages of the litigation lifecycle in order to meet legal obligations in a reasonable yet cost effective manner. The art of the science still lies in determining which technique to use, when.

Also, predictive coding relies heavily on the concept of giving you more of what you like. Brasso should seek to understand the nature of their own documents and conduct some form of high level review of the results – are the results comporting with their own expectations of their own documents given the issues in the case? Depending on the nature of the issues, the volume of documents collected, and the expected frequency of potentially responsive documents, the initial seed set of documents may need to be larger or perhaps can be made smaller.

LEGAL STANDARDS

This all sounds good and well, from a technical geek’s perspective. The practitioner’s question is typically (a) will it work in the real world, (b) is it defensible?

WILL IT WORK?

Predictive Coding appears to have great promise: “have a few attorneys review a seed set of documents and the system will do the rest.” At its base, it is supervised learning. The development of decision criteria based on a training set then applied to the rest of a set of information. In other words, predictions are made on the choices that would be made on the rest of the information.

It is not a single technology, as much as a technique. Variations of it form the underlying technology behind spam filters, which many people rely upon every day. Search terms can be used as a form of predictive coding. Unsupervised learning topics can be combined with document review as a form of predictive coding.

It is used to develop and or improve recommendations on Pandora (music), Netflix (movies), and Zite (news articles), for example. It is used by several software providers in the e-discovery space. But, it is used in many different ways. And the way in which it is used has a dramatic effect on the resulting advantages or disadvantages, as well as the decisions you need to take in how much effort to put into fine-tuning it.

As noted above in the definition section, sometimes the resulting predictions are relied upon with no further review. Sometimes the resulting predictions are used to prioritize the next step in review. Sometimes the resulting predictions are used to provide additional information to the user without forcing a particular answer.

IS IT DEFENSIBLE?

Predictive Coding can be many things. It can use neural networking, stochastic probabilistic distributions, Bayesian mathematics or even simple search terms. As mentioned above, it can be used to create a final decision, or to create a decision that will be confirmed by human review, and many variations in-between.

Therefore, ultimately, it is the process that is as important as the technology, and it is the results that matter. In the Victor Stanley case, Judge Paul Grimm, when describing search terms, laid out an either or scenario that is equally applicable here: either (a) the parties collaborate and agree to

the process and technology or (b) a single party creates a defensible process that includes thoughtful input on the front end, iterative improvement in the middle, and a test for over-inclusiveness and under-inclusiveness at the end.

Assuming you want to use technology and, given the recent very public discourse in Da Silva, you are concerned that you may not reach an agreement with your counter-party as to how to apply that technology, you should consider Judge Grimm's approach.

Consider your options carefully. Make thoughtful choices and provide thoughtful input. Thoughtful input can include the advice of seasoned counsel, a thorough review of the pleadings, discussions with fact witnesses and the early review of relevant documents.

Some technology tools are starting build in an automatic iterative process that helps them to improve through repeat sessions of review, prediction, testing and confirmation. This may include the manual review of documents that are known to be relevant, but missed by the technology (false negatives) and the review of documents that are known non-relevant, but flagged by the technology (false positives).

TESTING FOR PRECISION AND RECALL

Finally, it is now considered a best practice to review a sample of the residual documents to confirm that too many potentially responsive documents are not being left behind. In the case of using technology to make the final decision, final testing can also include using the technology to confirm that too many non-responsive documents are being included.

Looking at it from the requesting party's perspective can help. A requesting party would want you to know that responsive documents are not being left behind (an area at risk for under-inclusiveness) and that otherwise responsive documents are not inappropriately being flagged as privileged (an area at risk for over-inclusiveness).

Remember – we are applying technologies borrowed from the world at large and applying them to the niche of legal discovery. They are helpful, but there are differences to consider. For example, as a user of Netflix, Pandora and Zite, you may really like their service. You may feel that Netflix makes good movie suggestions for you. But, in the long run, you are not limited to those suggestions. You are not obligated to watch a movie that Netflix recommends and you are not restricted to only the recommendations that Netflix makes. This is an important distinction when using any type of predictive coding in a litigation context to make decisions that affect the quality and totality of documents being provided to counter parties.

Testing precision and recall is one way to test for both under-inclusiveness and over-inclusiveness. The more non-responsive documents you include, the lower the precision will measure. The more potentially responsive documents you miss, the lower the recall will measure.

Whether you measure precision and recall, which are well-defined terms, or chose another way to determine to your satisfaction that you have reasonably minimized the risk of under- and over-inclusiveness, it would be best to familiarize yourself with the science of quality control sampling. This can be a measured and reliable way to test your processes in a way that is easier to communicate to counter-parties and gain their confidence in your process.

In our hypothetical, Brasso engaged in a statistically measured sampling of their residual documents to develop an estimate of under-inclusiveness. Because Brasso reviewed each of the

documents suggested by the technology, Brasso would not need to conduct a sampled review for over-inclusiveness – they’ve already conducted an exhaustive one.

LEGAL STANDARD OF REASONABLENESS

All of the measuring and decision making sounds complicated. And it can be. Fortunately it doesn’t have to be perfect (and there are various academic studies that suggest it never will be perfect). Instead the standard of care is reasonableness as well as a sincere effort to abide by the Federal Rules of Civil Procedure, including F.R.C.P. 1, in applying the rules “to secure the speedy, just and inexpensive resolution of every action and proceeding.”

CASE LAW IN THIS AREA

Well, there isn’t any, yet. At least not much. There are two cases that are focused on the issues of predictive coding. These are the *Da Silva* matter in front of Judge Andrew Peck, and the *Kleen Products* matter in front of Judge Nan Nolan. These are two cases that present questions about the reliability and the applicability of predictive coding in a litigation context. Both are very new and both are unresolved.

Prior to these two matters, the majority of the case law is focused on the use of Search Terms, which is analogous in many ways to predictive coding. The prevailing opinion in this area is from Judge Paul Grimm in the *Victor Stanley* matter, which is referenced earlier in this paper, essentially laying out ground rules for how to conduct a robust and defensible process.

RESOURCES

Many parties feel that they don’t have the time, resources or experience to apply these technologies. For this reason, it is more common to see expert consultants and knowledgeable e-discovery counsel engaged in large matters.

Some of the leading judges in the area of e-discovery are taking an interest in helping parties understand and apply technology within document productions. Familiarity with opinions and presentations by Judge Lee Rosenthal, Mag. Judge Paul Grimm, Judge Shira Shindlein, Mag. Judge John Facciola, Mag. Judge Dave Waxse, Mag. Judge Nan Nolan, Mag. Judge Andrew Peck, and Mag. Judge Ron Hedges (ret.) can help.

There are several e-discovery conferences and think tanks that are providing education and thought leadership in this area, including the ABA Section of Litigation, The Sedona Conference, The Georgetown Advanced Institute for E-Discovery, the E-Discovery Institute and the Masters Conference.

SUMMARY

Predictive coding is a technique to apply decisions (human derived or machine derived) to a group of documents. There are several technologies that can be used (or combined) and several different workflows that rely solely, mostly or partially on the outcome. The workflow that you chose will be as significant as the technologies you incorporate and, in the long run, the defensibility of your choices will hinge on the reasonableness of your workflow (process) and the testing that was performed to confirm it. But, if you’re willing to follow a robust process and include quality control testing, you may be able to benefit from similar time and cost savings that are being claimed by others.