ICAIL 2011/DESI IV

Workshop on Setting Standards for Searching Electronically Stored Information In Discovery Proceedings

June 6, 2011

Thirteenth
International Conference
on
ARTIFICIAL INTELLIGENCE and LAW

ICAIL 2011

University of Pittsburgh School of Law, Pittsburgh PA

DESI IV Workshop Organizing Committee

Jason R. Baron, US National Archives and Records Administration, College Park MD
Laura Ellsworth, Jones Day, Pittsburgh PA
Dave Lewis, David D. Lewis Consulting, Chicago IL
Debra Logan, Gartner Research, London, UK
Douglas W. Oard, University of Maryland, College Park MD

TABLE OF CONTENTS

Research Papers

1. Thomas I. Barnett and Svetlana Godjevac, *Faster, Better, Cheaper Legal Document Review, Pipe Dream or Reality?*
2. Maura R. Grossman and Gordon V. Cormack, *Inconsistent Assessment of Responsiveness in E-Discovery: Difference of Opinion or Human Error?*
3. Richard T. Oehrle, *Retrospective and Prospective Statistical Sampling in Legal Discovery*

Position Papers

4. Steve Akers, Jennifer Keadle Mason and Peter L. Mansmann, *An Intelligent Approach to E-Discovery*
5. Susan A. Ardisson, W. Scott Ardisson and Decker, *bit-x-bit, LLC*
6. Cody Bennett, *A Perfect Storm for Pessimism: Converging Technologies, Cost and Standardization*
7. Bennett B. Borden, Monica McCarroll and Sam Strickland*, Why Document Review is Broken*
8. Macyl A. Burke, *Planning for Variation and E-Discovery Costs*
9. David van Dijk, Hans Henseler and Maarten de Rijke, *Semantic Search in E-Discovery*
10. Foster Gibbons, *Best Practices in Managed Document Review*
11. Chris Heckman, *Searches Without Borders*
12. Logan Herlinger and Jennifer Fiorentino, *The Discovery Process Should Account for Iterative Search Strategy*
13. Amanda Jones, *Adaptable Search Standards for Optimal Search Solutions*
14. Chris Knox and Scott Dawson, *ISO 9001: A Foundation for E-Discovery*
15. Sean M. McNee, Steve Antoch and Eddie O'Brien, *A Call for Processing and Search Standards in E-Discovery*
16. Eli Nelson, *A False Dichotomy of Relevance: The Difficulty of Evaluating the Accuracy of Discovery Review Methods Using Binary Notions of Relevance*
17. Christopher H. Paskach and Michael J. Carter, *Sampling – The Key to Process Validation*
18. Jeremy Pickens, John Tredennick and Bruce Kiefer, *Process Evaluation in eDiscovery as Awareness of Alternatives*
19. Venkat Rangan, *Discovery of Related Terms in a Corpus using Reflective Random Indexing*
20. Howard Sklar, *Using Built-in Sampling to Overcome Defensibility Concerns with Computer-Expedited Review*
21. Doug Stewart, *Application of Simple Random Sampling in eDiscovery*

Faster, *better*, *cheaper* legal document review, pipe dream or reality?

Using statistical sampling, quality control and predictive coding to improve accuracy and efficiency

Thomas I. Barnett and Svetlana Godjevac[1]

Iron Mountain

Abstract 1
Introduction 2
Background 3
Data Set and experiment 3
Data Set 3
Training 5
The Task 5
Coding Results 5
Analysis 6
Global Agreement Analysis 6
Pair-wise Analysis 8
Kappa 8
Other Industry Standards 9
Discussion 11
Recommendations 13
Conclusion 14
References 15
APPENDIX 16

Abstract

This paper examines coding applied by seven different review groups on the same set of twenty eight thousand documents. The results indicate that the level of agreement between the reviewer groups is much lower than might be suspected based on the general level of confidence on the part of the legal profession in the accuracy and consistency of document review by humans. Each document from a set of twenty eight thousand documents was reviewed for responsiveness, privilege and relevance to specific issues by seven independent review teams. Examination of the seven sets of coding tags for responsiveness revealed an inter-reviewer agreement of 43% for either responsive or non-responsive determinations. The agreement on the responsive determination alone was 9% and on the non-responsive determination was 34% of the total document family count. Pair-wise analysis of the seven groups of reviewers provided higher rates, however no pairing of the teams indicated that there is an unequivocally

[1] Thomas I. Barnett is the leader of the e-Discovery, records and information management consulting division of Iron Mountain, Inc.; Svetlana Godjevac is a senior consultant at Iron Mountain, Inc.

superior assessment of the dataset by any of the teams. This paper considers the ramifications of low agreement of human manual review in the legal domain and the need for industry benchmarks and standards. Suggestions are offered for improving the quality of human manual review using statistical quality control (QC) measures and machine-learning tools for pre-assessment and document categorization.

Introduction

In the world of technology assisted searching, analysis, review and coding of documents in litigation, review by human beings is typically viewed as the gold standard by which the accuracy and reliability of computer designations is measured. Similarly, humans are expected to be able to make judgments with computer-like accuracy and consistency across large sets of data. Expecting computer-like consistency from humans and expecting human-like reasoning from computers is bound to lead to disappointment all the way around. The level of quality of human review of a small number of documents by an expert reviewer familiar with the facts and issues in the matter is in fact a gold standard. But, the typical case involves review of large amounts of data by professional review teams not immersed in the subject matter of the case and the level of accuracy and consistency vary greatly. The levels of accuracy demanded of automated approaches to document classification are expected to confirm to the subject matter expert gold standard not the standard of the typical professional review team. The vast majority of data in legal document review is coded by professional review teams not by the subject matter experts. Thus, holding automated approaches to the gold standard that is barely, if ever, reached in the human review in actual matters creates an unreasonable and likely unachievable goal. This paper proposes that the comparisons be done on a level-playing field and that each approach, human and automated review, be applied to tasks to which they are best suited.

As more human reviewers are applied to the same set of data, the level of consistency and agreement predictably declines. This paper suggests that statistical sampling and statistical quality control is needed to establish a uniform framework from which to assess and compare human and automated review.

The tools used to search, analyze and make determinations about documents in a set of data need to be calibrated and guided by human understanding of the underlying facts and issues in the matter. For now at least, and with acknowledgement of the resounding victory by IBM's *Watson* on *Jeopardy!*, computers don't "understand" things in the way human beings do. Computers can execute vast amounts of simple binary calculations at speeds that are difficult to contemplate. Such calculations can be aggregated and structured in complex ways to mimic human analysis and decision making. But in the end, computers do exactly what they are told and are incapable of independent thought nor can they make decisions outside the scope of their programmatic instructions. Conversely, human beings do not blindly execute precise complex instructions at lightning speed in a predictable and measurable way as computers do. Human creativity and independent thought result in variability and unpredictability when attempting to make large numbers of fine distinctions. The independence and creativity that allows a person to make a novel observation or discovery is the flip side of the lack of the ability to make fast, mechanically precise consistent determinations about documents. This paper proposes considering a set of documents for review in a litigation as a continuum of relevance to a set of criteria rather than as a set of uniform discreet yes/no determinations. Under that model, the review process can be designed to play to the relative strengths of computer and human analysis. Within any typical set of data, certain documents will be clearly responsive. Others will be clearly non-responsive. The remaining documents can be characterized as having an ambiguous classification. Trying to get computers to accurately assess documents that humans find ambiguous is not effective—it plays to the computer's weakness. Computers should be utilized where they are strongest—quick, fast, accurate determinations of clear cut binary determinations. By contrast, for documents that are not clearly responsive or non-responsive, human judgment, creativity and flexibility is best suited to make the judgment calls. Based on this model, this

paper asserts that computers should be used to classify non-ambiguous documents while human reviewers should focus attention on documents whose classification is ambiguous.

This paper examines coding applied by seven different review groups on the same set of twenty eight thousand documents. The results indicate that the level of agreement between the reviewer groups is much lower than might be suspected based on the general level of confidence on the part of the legal profession in the accuracy and consistency of document review by humans (see Grossman and Cormack, 2011 for a similar position). However, a comparison to other industries, such as medical text coding for example, suggests that the legal industry is on a par with the results in other industries. This should not be surprising considering that both tasks are language-based tasks involving interpretation and translation of vast amounts of text into a single numeric code. This paper argues that the identified distribution of disagreements among human reviewers suggests that the nature of the task itself will never allow significant improvement in human review without disproportionate additional cost and time spend reviewing and cross checking document determinations. A proposed method to achieve higher consistency and accuracy lies in redistribution of the task between humans and computers. Computers should be allowed to jump-start the review, as they will easily recognize high-certainty sets, and humans should focus on ambiguous, middle of the scale sets, as only human analytical and inferential ability can successfully classify the documents of ambiguous classification.

Background

This experiment was originally conducted as a pilot by a company for the purpose of selecting a provider of document review services. The intent was to compare the document coding of five different document review providers against a control set of the same documents coded by outside counsel. The results of the six team review (five document review vendors and the outside counsel team) proved inconclusive to client in determining which provider to select. Subsequently, the client decided to assess the quality and accuracy of the providers' coding of the documents using the assessments of a different outside counsel who had reviewed the same set of documents. This second control group constituted the seventh set of human manual assessments for each document in this set. The additional control group's document coding determinations were ultimately not considered definitive and the pilot did not result in any clear "winner."

The analysis was performed on the final aggregate set of document coding from all review teams and does not assume that the coding of any one group is the ground truth. The client concluded that neither of the two control groups was able to provide coding that was of sufficient accuracy to be considered a gold standard. From the client's perspective, the experiment failed, as it was not possible to determine a winner among the document review service providers. Nevertheless for purposes of this analysis, the data provided a unique and valuable source of information for the eDiscovery industry and it is hoped that the results can be instructive in conducting comparisons of document review groups as well as creating quality control standards and workflow improvements for legal document review.

Data Set and experiment

Data Set

The reviewed document population for this experiment consisted of a sample of the electronically stored information (ESI) from six different custodians. The starting set contains 12,272 families comprised of 28,209 documents. Of the total 28,209 documents, most of the documents were emails and Microsoft Office application files. The basic data composition is represented in Figure 1. The most common family

unit[2] size was two. The majority of the corpus, 99%, consisted of families with no more than eight attachments. The family size frequencies are provided in Figure 2.

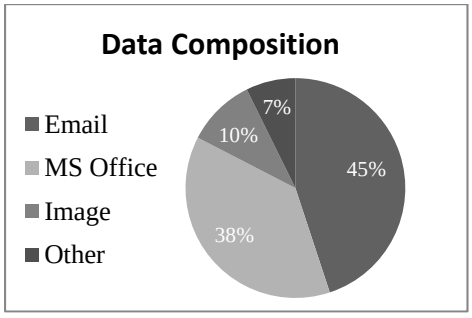


FIGURE 1- DATA COMPOSITION OF THE REVIEW SET

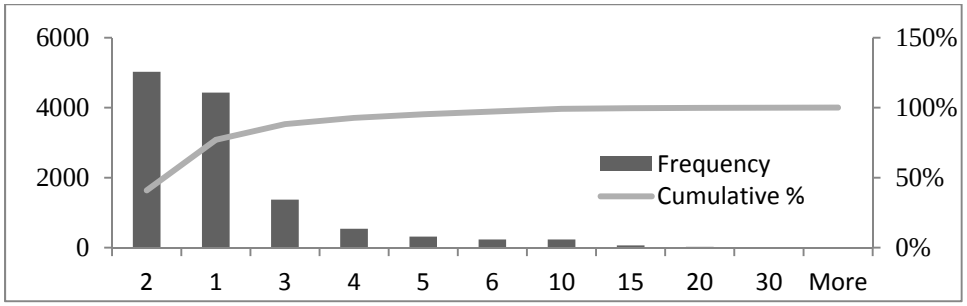


FIGURE 2 – FREQUENCY DISTRIBUTION OF FAMILY-UNIT SIZE – Most families consisted of two or one member.

Bin	Frequency	Cumulative %
2	5023	40.93%
1	4432	77.05%
3	1375	88.25%
4	542	92.67%
5	318	95.26%
6	235	97.17%
7-10	233	99.07%
11-15	66	99.61%
16-20	25	99.81%
21-30	13	99.92%
31 or More	10	100.00%

TABLE 1 – HISTOGRAM TABLE FOR THE FREQUENCY OF DISTRIBUTION OF SIZE OF FAMILY-UNITS

Due to errors in coding, the original set had to be cleaned up for the purpose of analysis. Forty-seven document families were excluded because at least one member has been coded "Technical Issue." Ninety five families were excluded because one or more members in the family were not coded consistently with the rest of the family. A summary of the data exclusion is presented in Table 2.

	ORIGINAL	EXCLUDED TECH ERRORS FAMILIES	EXCLUDED INCONSISTENT FAMILIES	CONSISTENT FAMILIES FINAL COUNT
Documents	28,209	205	350	27,654
Families	12,272	47	95	12,130

TABLE 2 – DATA SETS THAT WERE EXCLUDED FROM THE ORIGINAL SET AND THE FINAL SET COUNTS

[2] A "family unit" for purposes of this paper means an email and any associated attachments.

Reviewers
Seven reviewer groups were provided with access to the data for assessment. The review was conducted by groups of attorneys employed by five different legal document review providers and groups of litigators at two different law firms. Each group had a range of between six and seventeen attorneys who were provided access to the data.

Training
Each reviewer group received approximately three hours of subject matter training by the first law firm and the client. They were also provided with a review protocol, a coding manual, and an hour of training on the review platform. Each reviewer also received a binder with the review protocol, the official complaint, a list of acronyms and other subject matter materials necessary for document assessment. All but one team used the same hosted review platform which they accessed in a controlled environment during business hours. One group, group F, performed the review on their own platform, although there is no data to suggest that that influenced the document coding decisions.

The Task
The documents were arranged into batches of approximately 100 (keeping family units together). The batches were made up of randomly selected document families from the data set. The task involved reviewing and coding each document in the batch before the next batch could be requested. The coding tags included assessments for responsiveness, privilege, issue, and "hot" (significant) document designations. The assessments were made at the family unit level rather than by the individual component of a message unit. For example, if any member of the family was considered responsive, the entire family was coded responsive. Similarly, if any member of the responsive family was considered privileged, the entire family was tagged privileged. Each review team performed quality control checks according to their standard practice before providing the coded documents to the client.

Reviewers also had an option to tag documents for any technical problems, such as difficulty in viewing or errors in processing. Some of these errors prevented reviewers from making assessments for responsiveness and privilege. Consequently, due to the absence of coding for responsiveness, 205 documents were excluded from the overall agreement comparisons.

For purposes of analysis, responsiveness determinations were the sole focus. Unlike issue coding, these assessments are binary and all documents must be coded either responsive or non-responsive. Privilege determinations were not included because the privilege rates were very low, less than 1%, and were dependent on the responsive assessment (i.e., if a document was coded non-responsive, no determination would be made as to whether or not it was privileged).

Coding Results

The responsiveness rates among the seven review groups range from 23% to 54% of the total families. The difference spans 31% with a standard deviation of 0.11. The coding of each review group is presented in Table 3 below.

	Group						
Tag Count per Family	**A**	**B**	**C**	**D**	**E**	**F**	**G**
Non-Responsive	8279	5560	7641	9331	8842	6054	7316
Responsive	3851	6570	4489	2799	3288	6076	4814
Total	12130	12130	12130	12130	12130	12130	12130
Responsive Rate	**31.75%**	**54.16%**	**37.01%**	**23.08%**	**27.11%**	**50.09%**	**39.69%**

TABLE 3 – CODING COUNTS FOR EACH REVIEW TEAM

By definition, the global inter-reviewer agreement (the percentage of document coding all groups agree on) cannot exceed the lowest responsiveness rates found among all seven groups. In other words, the maximum rate of agreement cannot be higher than the sum of the lowest proportion of responsive tags among all the teams and the lowest proportion of non-responsive tags among all the teams (i.e., 23.08%+45.84%=68.92%).

Analysis

Two types of analyses were conducted: a global analysis of agreement, and a pair-wise agreement analysis. In the global analysis, the level of agreement between all reviewer groups was the focus. Sets of documents on which different teams agreed were identified: a set of documents for which all seven groups agreed, or 7/7, sets of documents for which six out of the seven agreed, or 6/7, five out of seven, 5/7, and four out of seven, 4/7. The remaining combinations are the inverse of these four. The pair-wise analysis was performed in two ways: an agreement expressed as a percent overlap between a pair of review teams and agreement expressed as Cohen's Kappa coefficient.

Global Agreement Analysis

The analyzed document set had 12,130 family units with a total of 27,654 documents. The set of families for which all seven groups agreed on responsiveness (either the responsive or non-responsive tag), is 5,233 family units, or 43.14% of the data set. Six groups agreed on 2,482 family units, or 20.46% of the data. Five groups agreed on 2,120 family units, or 17.48% of the data and four groups agreed on 2,295 family units, or 18.92% of the data. The agreement results are shown in Figure 3.

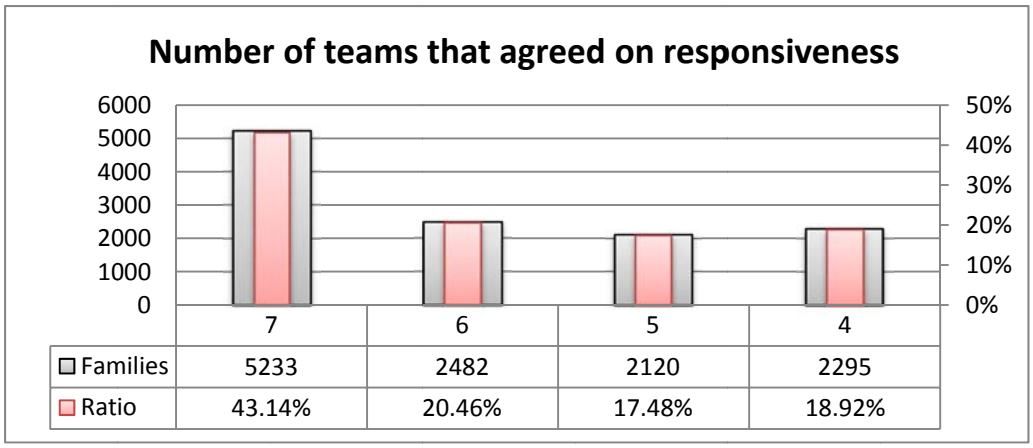


FIGURE 3 – REVIEWER AGREEMENTS ON RESPONSIVENESS
The chart shows the number of document families and the number of teams that tagged the documents the same way. For example, all seven teams coded 5233 families the same way.

The data in Figure 3 include agreements on both responsive and non-responsive determinations. Breaking down this agreement into its constituent parts and considering only the responsive tag (the non-responsive tag is a mirror image of the responsive tag) shows that the reviewers agreed more often on non-responsive than on the responsive tags. The distribution of the responsive tag agreement is provided in Figure 4.

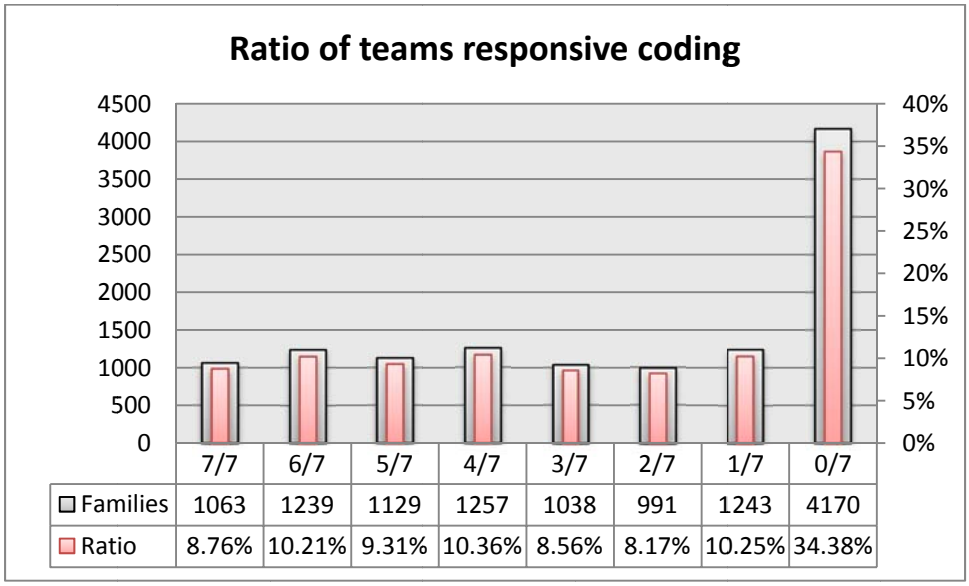


FIGURE 4 – DISTRIBUTION OF THE RESPONSIVE TAG ACROSS REVIEW GROUPS
The chart shows how many document families different number of teams' coded responsive. For example, seven review teams agreed on 1063 families being responsive; 6 review teams agreed on 1239 families being responsive etc. No group agreed that 4170 families were responsive, i.e., all seven groups coded this set as non-responsive.

All-groups agreement of 43.14%, shown in Figure 3, is the sum of the 8.76% document pool for which all seven teams said the document family was responsive and the 34.38% document pool for which all seven teams said the document family was non-responsive. The higher non-responsive agreement can be viewed as a result of the low responsive rate found in most groups coding. Two out of seven reviewer groups had responsive rates less than 30% (groups D and E, see Table 2).[3] With fewer documents coded responsive by two groups, the overall agreement for responsive would not be expected to be higher than the lowest responsive rate (Group D). Similar reasoning can be applied to the non-responsive rates.

If the seven groups applied the coding simply by guessing, what level agreement would be expected? With the seven review teams, a pure guessing approach would be the equivalent of seven coin tosses for each family of documents. Each coin toss is a binary decision, heads or tails, similar to document responsiveness tagging. The probability of having a chance agreement by having reviewers guess rather than apply analysis and reasoning among the seven groups is 1.56%, or 2/128.[4] The achieved 43% agreement is thus evidence of a decision and not a mere guess. However, the decision would require more specificity if it were to be executed in perfect accordance at a higher frequency than 43%. The next consideration was pair-wise comparisons[5] and the level of correlation among the group pairs.

[3] This is evidence that the actual, though unknown, number of responsive families, is much lower than the number of the actual, also unknown, number of non-responsive families.

[4] The probability for seven out of seven agreement on each responsive or non-responsive is $\left(\frac{1}{2}\right)^7$, which is $\left(\frac{1}{128}\right)$. Since aggregate agreement was computed, i.e., including agreement on both responsive and non-responsive those two probabilities are added to arrive at $\left(\frac{2}{128}\right)$.

[5] Pair-wise comparison is a common scientific method for calculating a relationship between a pair of results to determine which member of the pair is better or has a greater level of property that is under discussion.

Pair-wise Analysis

This analysis presents calculations of percent overlap (or agreement) between any two groups. The results are given in Table 4. Overlap is defined as the sum of all document families where two review teams agreed in responsiveness (responsive and non-responsive tag agreement) divided by the total number of document families they reviewed. The raw agreement values are shown in Table 9 in the Appendix. [6]

	A	B	C	D	E	F	G
A							
B	75.06%						
C	83.05%	75.01%					
D	74.51%	65.53%	72.20%				
E	79.91%	71.95%	76.69%	80.32%			
F	76.94%	84.90%	75.21%	68.17%	74.26%		
G	76.94%	75.23%	74.11%	67.39%	73.08%	77.20%	

TABLE 4 – PAIR-WISE AGREEMENTS

The table presents percent overlap of tagging assessments between a pair of review teams. For example, A and B teams tagging overlapped 75% of the time.

The highest overlap was achieved by groups A&C (83%) and B&F (85%). The lowest overlap was manifest between groups B&D (66%). The average overlap between group pairs is 75%. The group average aligns very closely with the results from a recent study by Roitblat et al.(2010) that compared agreement of pairs of manual review teams. Their comparison of manual review indicated that two different human review teams agreed with the original assessment at remarkably similar levels to the ones presented here. Their Team A agreed with the original review 75.58%, and Team B agreed with the original review 72.00%. So, results presented here replicate and reinforce the results presented in Roitblat et al. (2010). However, an earlier TREC study (Voorhees 2000) provided much lower agreement levels. In that study three different pairs of manual review teams had overlaps of 42.1%, 49.4% and 42.6%. It is not clear though, how the difference in ~30% agreement between the more recent studies and Voorhees' might be accounted for.

The average 75% coding overlap between two review teams suggests that even among the professional reviewers one in every four documents is not agreed upon. This result challenges the common assumption that there are discernable right and wrong determination for every document and that such a determination will be reached uniformly by different human reviewers.

Kappa

To further examine the level of agreement of responsiveness tagging between reviewer groups, Cohen's Kappa coefficient was computed. The Kappa coefficient is a measure of a level of agreement between two judges on a sorting of any number of items into a defined number of mutually exclusive categories. In our scenario, each review team is a judge and responsiveness tagging is a sorting into two mutually exclusive categories (responsive and non-responsive). Kappa coefficient values can range between 1 (complete agreement, or far more than expected by chance) to -1 (complete disagreement, or far less than expected by chance), with 0 being a neutral case, or as one would expect by pure chance. This coefficient is regarded as a better measure of agreement than percent-overlap because it eliminates the level of chance-agreement from its value. Landis and Koch (1977) propose the following interpretation of Kappa scores:

[6] Overlap presented in Table 4 was calculated from the values provided in Table 9. A and B teams agreed on 3698 document families being responsive and 5407 document families being non-responsive. Their coding then overlapped 75.06% ((3698+5407)/12130=0.7506).

0.01-0.20 – Slight agreement
0.21-0.40 – Fair agreement
0.41-0.60 – Moderate agreement
0.61-0.80 – Substantial agreement
0.81-0.99 – Almost perfect agreement

Kappa values for the seven review groups are presented in Table 5. Using the Landis and Koch interpretation scale for the Kappa scores, most of the team pairs, 13 of them, show moderate agreement. Their Kappa values range from 0.45 to 0.54. Two team pairs show substantial agreement, and six team pairs show fair agreement. The lowest score is 0.3402 (Groups B &D), and 0.6979 is the highest (Groups B & F). The Kappa values confirm the pair-wise analysis of percent-overlap for the groups: B&F exhibit the highest overlap and B&D the lowest on both analyses.

	A	B	C	D	E	F	G
A							
B	0.5159						
C	0.6255	0.5108					
D	0.3655	0.3402	0.3536				
E	0.5175	0.4597	0.4709	0.4776			
F	0.494	0.6979	0.5044	0.364	0.4857		
G	0.5013	0.5131	0.4528	0.4053	0.4053	0.5441	

TABLE 5 – KAPPA COEFFICIENT

The Kappa scores range [0.3402 - 0.6979] is similar to the one found by Wang & Soergel (2010) in their study of inter-rater agreement between two groups of human reviewers. Their experiment involved four law students as the LAW team and four library and information studies students, as the LIS team. The goal of their experiment was to test whether the legal background affects the quality of document review. The Kappa mean scores within the LAW team, within the LIS team and across the two teams show remarkably similar ranges: (a) within LAW [0.38 – 0.69], (b) within LIS [0.30 – 0.54] and (c) across LAW and LIS [0.47 – 0.61]. The range of the Kappa coefficient for Wang and Soergel's LAW group closely parallels the range reported here for the seven review teams.

The Kappa coefficient analysis further confirms that humans reviewing the same documents frequently disagree. As discussed below, this fact suggests that greater focus on quality control is warranted.

Other Industry Standards

In order to put results presented here into a broader context, a short overview of similar tasks in other domains is presented. There are a variety of applications that require translation of natural language into other systems, whether other natural languages or man-made systems. Document review coding is an example of a man-made system that requires a translation from document text into review codes. Tasks of this nature could theoretically be automated if explicit sets of rules could accurately be defined in advance. For tasks that involve natural language, the number of explicit rules is too numerous to be able to be defined in advance. One solution to this problem is machine learning. Machine learning is a substitute for pre-defined set of rules. In the absence of explicit rules, a machine learning program uses input from a training set and "learns" how to apply it in situations that are similar to the ones in the training set. Machine learning is used in search engines, natural language processing, detecting credit card fraud, stock market analysis, handwriting recognition, game playing, medicine, and many others areas.

The training phase of machine learning requires high quality human input, where the high level of accuracy is confirmed through agreement with multiple human experts on the same task.

The medical industry has been faced with the challenge of coding millions of records for medical diagnosis, billing and insurance purposes, among others. In the domain of patient records, a medical diagnosis is required to be translated into a billing code. The billing codes are based on the classification provided by the World Health Organization in the International Classification of Diseases (ICD). The process of human coding of medical diagnoses is challenged by the existence of thousands of possible codes, which is both time-consuming and error-prone. To alleviate the burden and improve consistency of human coding, a number of machine learning systems for classifying text using natural language processing have been designed and implemented in the medical industry. The "training" of the system, using a set of documents that have been coded by highly trained human ICD coding experts is critical to the accuracy of all of the ICD automated coding systems.

The application of ICD codes for medical diagnosis is in many ways similar to legal document review. Both involve reading and understanding natural language texts (or listening to audio files) and applying a code as an output of the process. The ICD codes are directly parallel with issue coding in legal document review in that a number of possibilities per document are open for assignment. The interpretation of natural language (verbal encoding of someone else's intentions) is at the core of the process in both tasks. Responsive and privilege binary distinctions are a simpler form of coding than relevance to a specific issue in a lawsuit as the number of possibilities are reduced to two. So, the agreement results achieved in responsiveness tagging are expected to be higher than agreements on issue tagging in legal review or ICD coding in medical review due to the smaller number of choices a reviewer/coder is faced with.

The literature on training and automation of the ICD coding assignment and other systems for classification of medical information, such as SNOMED (Systematized Nomenclature of Medicine), is vast. Kappa is often used as a measure of inter-reviewer agreement and for comparison of automated system against human review, the most commonly used metric is the harmonic mean of precision and recall, or the F-score.[7] This score can only be computed if precision and recall can be computed. Having a gold standard is the key to all machine learning systems as well as the evaluation metrics. If the "true" answer is unavailable, the system is unable to learn.[8] Some examples of results in the medical domain are provided below.

Uzuner et al. (2008) measured inter-annotator agreement of the patient's smoking status based on the hospital discharge summary. The annotators were two pulmonologists who provided annotations relying on the explicit text in the summary as well as their understanding of the same text. The metric shown in Table 6 is the Kappa coefficient. The intuitive judgment values are the most directly comparable to the document review assessments as they rely on human ability for interpretation. These scores are similar to the ones reported here for attorney teams. The overall range is wider with the highest score in the "almost perfect" category.

[7] The F-score is computed as 2*P*R/(P+R), where P is precision and R is recall. Precision is a metric that quantifies how many of the retrieved documents are correct and precision is a metric that quantifies how many correct documents were missed. In order to calculate these values, the number of correct documents must be known. The set of correct documents is what is referred to as the "gold standard."

[8] Human intelligence, although incomparably more flexible and dynamic in comparison to a machine, is also dependent on the "system updates", or the feedback loop for arriving at the truth. Quality control checks of a sample of documents being reviewed often serve to provide feedback to the reviewers on the accuracy of their coding choices so that they can make course-corrections going forward. This process is an important calibration tool in manual review.

Agreement	Textual Judgment	Intuitive Judgment
Observed	0.93	0.73
Specific (Past Smoker)	0.85	0.56
Specific (Current Smoker)	0.72	0.44
Specific (Smoker)	0.40	0.30
Specific (Non-Smoker)	0.95	0.60
Specific (Unknown)	0.98	0.84

TABLE 6- KAPPA COEFFICIENTS FOR INTER-ANNOTATOR AGREEMENT FOR PATIENT'S SMOKING STATUS
From Uzuner et al. 2008 study on patient smoking status from medical discharge summaries. The study shows the kappa scores for assessments based on explicit text and interpretive judgments based on human understanding.

Table 7, below, shows pair-wise comparison of inter-reviewer agreement using the F-measure, for three human annotators for ICD-9-CM codes applied to radiology reports on a test set (unseen data). The F-scores of the training set were approximately 2 points higher in each case. This higher measure is as one would expect, as the training set is the set that they've seen prior to evaluation.

	A1	A2	A3
A1		73.97	65.61
A2	73.97		70.89
A3	65.61	70.89	

TABLE 7- INTER-ANNOTATOR AGREEMENT ON ICD-9-CM CODING OF RADIOLOGY REPORTS (Richard Farkas And Gyorgy Szarvas, 2008)

Crammer et al. (2007) study of inter-annotator agreement for ICD-9-CM coding of free text radiology reports, also using three human coders, the average F-measure of 74.85 (with standard deviation of 0.06). Resnik et al. (2006) provide measures of inter-annotator agreement on task involving code application for for ICD-9-CM and CPT (Current Procedural Terminology) on a random sample of 720 radiology notes from a single week from a large teaching hospital. Their evaluations show averages for all annotators. They've used a proportion measure for ICD. Their results are provided in Table 8.

	ICD
Intra-coder agreement	64%
Inter-coder agreement	47%

TABLE 8 – INTER AND INTRA CODER AGREEMENT ON ICD CODE ASSIGNMENTS (Resnik et al. 2006)

In all of the radiology coding tasks presented above, the inter-reviewer agreement is not dramatically different from the agreements found in this study of legal document review. Given that similarity of tasks, this suggests that manual (human) review of discovery documents should not be expected to improve significantly unless additional means are used to help better allocate time for human review of more complex documents that need to be assessed with more attention.

One common thread to the medical studies and the studies on legal document review, whether by humans or machines, referenced here is the fact that none of them show results approaching full agreement or high retrieval (measured by the F-score). Both fields appear to be at the same level of advancement when it comes to coping with the inherent ambiguity of human language.

Discussion

Results and their implication

The global agreement calculations show that reviewers unanimously agreed on nearly half the documents, or 43%. This set of documents can be termed a high certainty set. On the other roughly half of the

documents, the reviewers had varying degrees of certainty, 6/7, 5/7, and 4/7. This distribution of varying degrees of collective uncertainty can be viewed as a consequence of the "translation" reviewers had to make in order to force a simple yes/no determination onto intrinsically subjective nonlinear data. In other words, the perspective of multiple review groups reviewing the same set of documents rather than a single review team provides support for the intuitive understanding that documents have varying degrees of relevance. When reviewers are asked to code documents either responsive or non-responsive, they are essentially being asked to translate a continuum of degrees of responsiveness into a threshold that will create a single artificial boundary for a yes/no determination. Where this boundary lies is subject to interpretation. The subject matter training the reviewers receive at the beginning of a review is supposed to train them to find this boundary uniformly at the same place every time. However, in reality, each reviewer (and consequently each group) arrives at a different threshold that defines that boundary. Quality control is needed to moderate the understanding of the boundary placement throughout the review. The level of QC needed to guarantee that this boundary is perfectly calibrated and aligned for all reviewers is not practical in terms of time and cost in the context of legal document review.

Part of the quality of control process in the context of document review is evaluation of performance. The most effective means of evaluating quality of performance is to use a quantifiable system. Often used steps for quantifiable evaluation of language-based tasks are:

a) comparison to a gold standard
b) inter-coder agreement (consistency across multiple reviewers)
c) intra-coder agreement (consistency within the same reviewer)

This study of agreement only focused on inter-coder agreement. Access to a gold standard was unavailable and inter-reviewer consistency either at the group-level or reviewer-level would require more complex computations such as creating document sub-groupings based on content similarity and assessing consistency of coding within each subgroup within a reviewer, within a reviewer team and across all reviewer teams.

Comparison of inter-reviewer agreement from these seven groups to the quoted radiology annotators shows that the legal review groups are on a par with the medical profession. The ICD proportion for inter-coder agreement was 47% (Table 8). This value is directly comparable to the average value of 75%, calculated in Table 5 for legal document review. The comparison of these two values gives legal review a superior grade. The comparison analysis, however, must acknowledge that the ICD coders use thousands of codes, rather the just two (i.e., responsive or non-responsive), as do the legal document reviewers and thus the probability of agreement is reduced by the larger number of possible choices.

If it is assumed that the set of varying degrees of certainty (the sets where agreements were 6/7, 5/7, and 4/7) and the sets outside of agreement (intersections) in the pair-wise comparisons are the sets that contain errors, the nature of these errors and the cost associated with them needs to be considered.

Error types and their cost

Errors are divided into two types:

- False positives (Type I error) – documents coded responsive, but are actually non-responsive.
- False negatives (Type II error) – documents coded non-responsive, but are actually responsive.

False positives are typically caught by QC and/or additional review passes. This is because the set of responsive documents is usually further reviewed either for assessment/confirmation of privilege, privilege type or redaction. Errors of this type, Type I, are usually more costly for the client in the field of legal document review, because these types of errors may result in waiver of privilege or revealing potentially damaging information to the opposing side.

False negatives and the degree of their presence in the non-responsive set usually remain undiscovered, unless active measures are taken to identify them such as re-review or inferential statistics through sampling.. This type of error is often neglected as it is less costly from the perspective of the risk of unintentional information exposure. However, if detected by the opposing side, it could lead to sanctions for withholding relevant information.

In this study, an assumption was made that the gold standard for this set was not available, However, if the set of 7/7 agreements for responsive and non-responsive were to be used as the gold standard, the calculation based on this gold standard would be biased in favor of the groups who made conservative judgments on responsiveness. So, this evaluation cannot be used as a measure of quality of the review groups, although it could be used as a way of measuring the cost of error for the client.

Recommendations

Sharing the work

The distribution of partial agreements, viewed as a continuum of degrees of certainty, is analogous to the predictive coding systems whose output is a probability score for each document, rather than a binary decision on category membership. If human review manifests a continuum of certainty levels with respect to relevancy judgment anyway, why not then share the task of review with the predictive coding systems which automatically output degrees of certainty?

Sharing the task does not mean fully delegating, but rather incorporating predictive coding technologies to aid human document review by using computer software to segregate the high-certainty sets (the high probabilities and the low probabilities for category membership, or the 7/7 and 0/7 agreements in this study) and allow human experts to focus on the middle range probabilities (the 6/7, 5/7, and 4/7 in this study). The high certainty sets are the easy calls to make as they are more clear-cut and so they should be delegated to the low cost (computer) labor. The difficult decisions are the decisions that require human intelligence for disambiguation as well as strong subject matter expertise.

The generated probabilities can also speed up the review of the middle of the scale sets. Resnik et al. (2006) show that computer assisted workflow improves human scores by 6% in ICD coding. This improvement in speed may come with a bias, however, and so, it should be considered carefully. They note that:

> *"Post hoc reviews can overestimate levels of agreement when complex or subjective judgments are involved, since it is more likely that a reviewer will approve of a choice than it is that they would have made exactly the same choice independently"*

Whether predictive coding should be revealed to the reviewers for the middle of the scale sets is a decision that will require determination on a case-by-case basis.

Feedback

Feedback is essential for any learning environment. Legal document review is a business process that starts anew with each case. The task begins typically after no more than a day of training, if that. Due to the high costs of document review by attorneys, the learning phase is becoming shorter and shorter and the expectation is that even very complex subject matters can be absorbed in short time frames. Unfortunately, that assumption is to the detriment of the depth of expertise reviewers can attain and consequently the quality of the review. The actual subject matter experts rarely review documents and thus the true gold standard is an illusion. To improve the quality of review, continuous dynamic updates of expert judgments provided to the reviewers are critical. If reviewers receive feedback about the

accuracy of their work promptly, fewer errors will ensue. This result will minimize the need for recoding after quality control checks are performed as fewer errors should be present.

Statistical QC

Current legal document review practices rely more often on judgmental sampling as a QC procedure than on statistical sampling. Although judgmental sampling has value in the QC process, it also has deficiencies. The key detriment is the inability to apply inferences to the larger set. So, while judgmental sampling may reveal errors, there is no way of estimating if the types of errors the QC team didn't consider are present and the degree to which they may be present in the population as a whole. For example, because judgmental sampling deals with the known risks the searches target known "keywords" to create samples for QC. The end result is that unanticipated uses of language to describe the high-risk activities at the core of review will remain undetected. Implementing statistical sampling for the QC process would allow document review to provide quantifiable metrics on the quality of the output and it would also create a higher chance of finding unanticipated references that may inform new searches and require document recoding.

As predictive coding is becoming a more widely available offering in the practice of legal document review, it is essential that the double standard that seems to be applied to this programmatic approach as compared to the standards for human review be addressed. Clients uniformly require that predictive coding come with 95%-99% accuracy. This level of accuracy for the machine is expected because the assumption is that human review is in the 100% range of accuracy (for a similar discussion see Grossman and Cormack 2011). There are at least two problems with this reasoning. First, no research was uncovered that suggests that human accuracy level ever approaches 100% accuracy. Second, it seems that this unsupported assumption is also tacitly known to be false. Either way, the predictive coding should be welcomed by the legal community and judged by the same, not higher, standards than manual review. In order to provide the ground for comparison and equivalent standards of quality, manual review should incorporate statistical QC into its workflow as only with this type of quality check can measures of accuracy, such as precision and recall, be calculated.

Conclusion

Document review for litigation discovery is demanding, time-consuming, expensive and risky. It requires both the ability to perform routine repetitive tasks in an accurate and timely manner as well as the ability to apply human judgment, reasoning and making fine distinctions about complex matters. And the faulty decisions can have tremendous legal and financial consequences. Neither humans nor computers are perfectly suited to accomplish these diverse tasks. The recommended approach to achieve greater accuracy and efficiency is to allocate tasks between humans and computers that play to their respective strengths rather than to their respective weaknesses. Computers perform high speed, repetitive tasks far more efficiently than humans. But computers have no ability to use reason, creativity or judgment beyond the predefined rule sets that are used to program them. Large sets of documents subject to review in litigation contain a continuum of responsiveness. That is, there are some documents that are clearly responsive, some that are clearly non-responsive and the remainder are somewhere in between. Efficiency and accuracy in legal document review can be improved by allocating computer assisted sorting and categorization processes to the high certainty ends of the continuum while human reviewers focus their time and attention using their uniquely human analytical and inferential ability classifying the ambiguous documents.

References

Richard Farkas and Gyorgy Szarvas, "Automatic construction of rule-based ICD-9-CM coding systems", BMC Bioniformatics 2008, 9.

Koby Crammer and Mark Dredze and Kuzman Ganchev and Partha Partim Talukdar, "Automatic Code Assignment To Medical Text", BioNLP '07 Proceedings of the Workshop on BioNLP 2007: Bilogical, Translational, and Clinical Language Processing 2007.

Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11 (2011), http://jolt.richmond.edu/v17i3/article11.pdf

Landis, J.R. and Koch, G. G., "The measurement of observer agreement for categorical data", *Biometrics* 1977, 33.

Philip Resnik, Micahel Niv, Micahel Nossal, Gregory Schnitzer, Jean Stoner, Andrew Kapit and Richard Toren, 2006, "Using Intrinsic and Extrinsic Metrics to Evaluate Accuracy and Facilitation in Computer-assisted Coding", *Perspectives in Health Information Management* Computer Assisted Coding Conference Proceedings; Fall 2006.

Roitblat, h. L., Kershaw, A., and Oot, P. "Document categorization inlegal electronic discovery: Computer classification vs. manual review", Journal of the American Society for Information Science and Technology 61 (2010), 70–80.

Özlem Uzuner, PhD, [a b] Ira Goldstein, MBA, [a] Yuan Luo, MS, [a] and Isaac Kohane, MD, PhD [c], "Identifying Patient Smoking Status from Medical Discharge", Journal of the American Medical Informatics Association. 2008 Jan-Feb; 15(1): 14-24.

Vooerhees, Ellen M., "Variations in relevance judgments and the measurement of retrieval effectiveness", Information Processing & Management 36, 5 (2000), 697–716

Wang, Jianqiang and Dagobert Soergel, "A User Study of Relevance Judgments for E-Discovery", ASIST 2010, October 22-27, 2010, Pittsburgh, PA.

APPENDIX

B \ A	R	NR	Total
R	3698	2872	6570
NR	153	5407	5560
Total	3851	8279	12130

C \ A	R	NR	Total
R	3142	1347	4489
NR	709	6932	7641
Total	3851	8279	12130

C \ B	R	NR	Total
R	4014	475	4489
NR	2556	5085	7641
Total	6570	5560	12130

D \ A	R	NR	Total
R	1779	1020	2799
NR	2072	7259	9331
Total	3851	8279	12130

D \ B	R	NR	Total
R	2594	205	2799
NR	3976	5355	9331
Total	6570	5560	12130

D \ C	R	NR	Total
R	1958	841	2799
NR	2531	6800	9331
Total	4489	7641	12130

E \ A	R	NR	Total
R	2351	937	3288
NR	1500	7342	8842
Total	3851	8279	12130

E \ B	R	NR	Total
R	3228	60	3288
NR	3342	5500	8842
Total	6570	5560	12130

E \ C	R	NR	Total
R	2475	813	3288
NR	2014	6828	8842
Total	4489	7641	12130

E \ D	R	NR	Total
R	1850	1438	3288
NR	949	7893	8842
Total	2799	9331	12130

F \ A	R	NR	Total
R	3428	2648	6076
NR	423	5631	6054
Total	3851	8279	12130

F \ B	R	NR	Total
R	5407	669	6076
NR	1163	4891	6054
Total	6570	5560	12130

F \ C	R	NR	Total
R	3779	2297	6076
NR	710	5344	6054
Total	4489	7641	12130

F \ D	R	NR	Total
R	2507	3569	6076
NR	292	5762	6054
Total	2799	9331	12130

F \ E	R	NR	Total
R	3121	2955	6076
NR	167	5887	6054
Total	3288	8842	12130

G \ A	R	NR	Total
R	2934	1880	4814
NR	917	6399	7316
Total	3851	8279	12130

G \ B	R	NR	Total
R	4190	624	4814
NR	2380	4936	7316
Total	6570	5560	12130

G \ C	R	NR	Total
R	3081	1733	4814
NR	1408	5908	7316
Total	4489	7641	12130

G \ D	R	NR	Total
R	1829	2985	4814
NR	970	6346	7316
Total	2799	9331	12130

G \ E	R	NR	Total
R	2418	2396	4814
NR	870	6446	7316
Total	3288	8842	12130

G \ F	R	NR	Total
R	4062	752	4814
NR	2014	5302	7316
Total	6076	6054	12130

TABLE 9 – THE CONTINGENT RELATIONS BETWEEN THE RESPONSIVENESS CODING OF SEVEN REVIEW TEAMS

Research paper at *DESI IV: The ICAIL 2011 Workshop on Setting Standards for Searching Electronically Stored Information in Discovery Proceedings*, June 6, 2011, University of Pittsburgh, Pittsburgh, PA, USA.
http://www.umiacs.umd.edu/~oard/desi4/

Inconsistent Assessment of Responsiveness in E-Discovery: Difference of Opinion or Human Error?

Maura R. Grossman, J.D., Ph.D.
Wachtell, Lipton, Rosen & Katz[1]

Gordon V. Cormack, Ph.D.
University of Waterloo

1 Introduction

In responding to a request for production in civil litigation, the goal is generally to produce, as nearly as practicable, *all* and *only* the non-privileged documents that are *responsive* to the request.[2] *Recall* – the proportion of responsive documents that are produced – and *precision* – the proportion of produced documents that are responsive – quantify how nearly *all* of and o*nly s*uch responsive, non-privileged documents are produced [2, pp 67-68].

The traditional approach to measuring recall and precision consists of constructing a *gold standard* that identifies the set of documents that are responsive to the request. If the gold standard is complete and correct, it is a simple matter to compute recall and precision by comparing the production set to the gold standard. Construction of the gold standard typically relies on human assessment, where a reviewer or team of reviewers examines each document, and codes it as responsive or not [2, pp 73-75].

It is well known that any two reviewers will often disagree as to the responsiveness of particular documents; that is, one will code a document as responsive, while the other will code the same document as non-responsive [1, 3, 5, 8, 9, 10]. Does such disagreement indicate that responsiveness is ill-defined, or does it indicate that reviewers are sometimes mistaken in their assessments? If responsiveness is ill-defined, can there be such a thing as an accurate gold standard, or accurate measurements of recall and precision? Answering this question in the negative might call into question the ability to measure, and thus certify, the accuracy of a response to a production request. If, on the other hand, responsiveness is well-defined, might there be ways to measure and thereby correct for reviewer error, yielding a better gold standard, and therefore, more accurate measurements of recall and precision?

This study provides a qualitative analysis of the cases of disagreement on responsiveness determinations rendered during the course of constructing the gold standard

[1] The views expressed herein are solely those of the author and should not be attributed to her firm or its clients.

[2] *See* Fed. R. Civ. P. 26(b) & (g), 34(a), and 37(a)(4).

for the TREC 2009 Legal Track Interactive task ("TREC 2009") [7]. For each disagreement, we examined the document in question, and made our own determination of whether the document was "clearly responsive," "clearly non-responsive," or "arguable," meaning that it could reasonably be construed as either responsive or not, given the production request and operative assessment guidelines.

2 Prediction

Our objective was to test two competing hypotheses:

> **Hypothesis 1:** *Assessor disagreement is largely due to ambiguity or inconsistency in applying the criteria for responsiveness to particular documents.*
>
> **Hypothesis 2:** *Assessor disagreement is largely due to human error.*

Hypothesis 1 and Hypothesis 2 are mutually incompatible; evidence refuting Hypothesis 1 supports Hypothesis 2, and vice versa.

To test the validity of the two hypotheses, we constructed an experiment in which, prior to the experiment, the two hypotheses were used to predict the outcome. An observed result consistent with one hypothesis and inconsistent with the other would provide evidence supporting the former and refuting the latter.

In particular, Hypothesis 1 predicted that if we examined a document about whose responsiveness assessors disagreed, it would generally be difficult to determine whether or not the document was responsive; that is, it would usually be possible to construct a reasonable argument that the document was either responsive or non-responsive. On the other hand, Hypothesis 2 predicted that it would generally be clear whether or not the document was responsive; it would usually be possible to construct a reasonable argument that the document was responsive, or that the document was non-responsive, but not both.

At the outset, we conjectured that the results of our experiment would more likely support Hypothesis 1.

3 TREC Adjudicated Assessments

The TREC 2009 Legal Track Interactive Task used a two-pass adjudicated review process to construct the gold standard [7]. In the first pass, law students or contract attorneys assessed a sample of documents for each of seven production requests – "topics," in TREC parlance – coding each document in the sample as responsive or not. TREC 2009 participants were invited to appeal any of the assessor coding decisions with which they disagreed, and the Topic Authority (or "TA") – a senior lawyer tasked with defining responsiveness – was asked to make a final determination as to whether the appealed document was responsive or not. The gold standard considered a document to be *responsive* if the first-pass assessor coded it as responsive and that decision was not appealed, the first-pass assessor coded it as responsive and that decision was upheld by the Topic Authority, or the first-pass assessor coded it as non-responsive and

Topic	First-Pass Assessment	Assessed	Appealed	Success	% Success
201	Responsive	603	374	363	97%
201	Non-responsive	5,605	123	101	82%
202	Responsive	1,743	167	115	68%
202	Non-responsive	5,462	541	469	86%
203	Responsive	131	74	69	93%
203	Non-responsive	5,296	209	186	88%
204	Responsive	105	59	50	84%
204	Non-responsive	7,024	207	169	81%
205	Responsive	1,631	889	882	99%
205	Non-responsive	4,289	78	50	64%
206	Responsive	235	52	50	96%
206	Non-responsive	6,860	0	0	–
207	Responsive	938	43	23	53%
207	Non-responsive	7,377	154	125	81%
All	Responsive	5,386	1,658	1,552	93%
All	Non-responsive	41,913	1,312	1,100	83%

Table 1: Number of documents assessed, appealed, and the success rates of appeals for the TREC 2009 Legal Track Interactive Task, categorized by topic and first-pass assessment.

that decision was overturned by the Topic Authority. The gold standard considered a document to be *non-responsive* if the first-pass assessor coded it as non-responsive and that decision was not appealed, the first-pass assessor coded it as non-responsive and that decision was upheld by the Topic Authority, or the first-pass assessor coded it as responsive and the decision was overturned by the Topic Authority.

A gold standard was created for each of the seven topics.[3] A total of 49,285 documents – about 7,000 per topic – were assessed for the first-pass review. A total of 2,976 documents (5%) were appealed and therefore adjudicated by the Topic Authority. Of those appeals, 2,652 (89%) were successful; that is, the Topic Authority disagreed with the first-pass assessment 89% of the time. A breakdown of the number of documents appealed per topic, and the outcome of those appeals, appears in Table 1.[4]

4 Post-Hoc Assessment

We performed a qualitative, post-hoc assessment on a sample of the successfully appealed documents from each category represented in Table 1; that is, the documents where the TREC 2009 first-pass assessor and Topic Authority disagreed. Where 50 or more documents were successfully appealed, we selected a random sample of 50.

[3] The gold standard and evaluation tools are available at `http://trec.nist.gov/data/legal09.html`.

[4] The pertinent documents may be identified by comparing files `qrels_doc_pre_all.txt` and `qrels_doc_post_all.txt` in `http://trec.nist.gov/data/legal/09/evalInt09.zip`.

Topic	TA Opinion	TA Correct	Arguable	TA Incorrect
201	Responsive	74%	20%	6%
201	Non-responsive	94%	2%	4%
202	Responsive	96%	2%	2%
202	Non-responsive	96%	0%	4%
203	Responsive	94%	2%	4%
203	Non-responsive	82%	4%	14%
204	Responsive	90%	10%	0%
204	Non-responsive	90%	8%	2%
205	Responsive	100%	0%	0%
205	Non-responsive	82%	4%	14%
206	Responsive	–	–	–
206	Non-responsive	96%	2%	2%
207	Responsive	74%	12%	14%
207	Non-responsive	70%	0%	28%
All	Responsive	88% (84–91%)	8% (5–11%)	4% (2–7%)
All	Non-responsive	89% (85–92%)	3% (2–6%)	8% (5–12%)

Table 2: Post-hoc assessment of documents whose first pass responsiveness assessment was overturned by the Topic Authority in the TREC 2009 Legal Track Interactive Task. The columns indicate the topic number, the TA's assessment, the proportion of documents for which the authors believe the TA was clearly correct, the proportion of documents for which the authors believe the correct assessment is arguable, and the proportion of documents for which the authors believe the TA was clearly incorrect. The final two rows give these proportions over all topics, with 95% binomial confidence intervals.

Doc. Id.	TA Opinion	Post-Hoc Assessment	TA Reconsideration
0.7.47.1151420	Responsive	Arguable	TA Incorrect
0.7.47.1310694	Responsive	Arguable	TA Incorrect
0.7.47.272751	Responsive	TA Incorrect	Arguable
0.7.6.180557	Responsive	Arguable	TA Correct
0.7.6.252211	Responsive	Arguable	TA Incorrect
0.7.47.1082536.1	Non-responsive	Arguable	TA Correct
0.7.47.14687.1	Non-responsive	Arguable	Arguable
0.7.47.758281	Non-responsive	Arguable	TA Correct
0.7.6.707917.2	Non-responsive	Arguable	TA Correct
0.7.6.731168	Non-responsive	Arguable	TA Correct

Table 3: Blind reconsideration of adjudication decisions for Topic 204 by the Topic Authority (Grossman) that were contradicted or deemed arguable by the post-hoc reviewer (Cormack). The columns represent the TREC document identifier for each of the ten documents, the opinion rendered by the TA during the TREC 2009 adjudication process, the opinion rendered by the post-hoc reviewer, and the *de novo* opinion of the same Topic Authority for the purposes of this study.

Where fewer than 50 documents were successfully appealed, we selected all of the appealed documents.

We used the plain-text version of the TREC 2009 Legal Track Interactive Track corpus, downloaded by one of the authors while participating in TREC 2009 [4], and redistributed for use at TREC 2010.[5] One of the authors of this study examined every document, in every sample, and coded each as "responsive," "non-responsive," or "arguable," based on the content of the document, the production request, and the written assessment guidelines composed for TREC 2009 by each Topic Authority. We coded a document as "responsive" if we believed there was no reasonable argument that the document fell outside the definition of responsiveness dictated by the production request and guidelines. Similarly, we coded a document as "non-responsive" if we believed there was no reasonable argument that the document should have been identified as responsive to the production request. Finally, we coded the document as "arguable" if we believed that informed, reasonable people might disagree about whether or not the document met the criteria specified by the production request and guidelines.

Table 2 shows the agreement of our post-hoc assessment with the TREC 2009 Topic Authority's assessment on appeal, categorized by topic and by the TA's assessment of responsiveness. Each row shows the TA opinion (which is necessarily the opposite of the first-pass opinion), the percentage of post-hoc assessments for which we believe that the only reasonable coding was that rendered by the TA, the percentage of post-hoc assessments for which we believe that either coding would be reasonable, and the percentage of post-hoc assessments for which we believe that the only reasonable coding contradicts the one that was made by the TA.

5 Topic Authority Reconsideration

One of the authors (Grossman) was the Topic Authority for Topic 204 at TREC 2009. The other author (Cormack) conducted the post-hoc assessment for Topic 204. The post-hoc assessment clearly disagreed with the Topic Authority in only one case, and was "arguable" in nine other cases. The ten documents were presented to the TA for *de novo* reconsideration, in random order, with no indication as to how they had been previously coded. For this reconsideration effort, the TA used the same three categories as for the post-hoc assessment: "responsive," "non-responsive," or "arguable."[6] Table 3 shows the results of the TA's reconsideration of the ten documents.

6 Document Exemplars

Table 4 lists the production requests for the seven TREC topics. Based on the production request and his or her legal judgement, each Topic Authority prepared a set

[5] Available at `http://plg1.uwaterloo.ca/~gvcormac/treclegal09/`.

[6] Note that when the TA adjudicated documents as part of TREC 2009, she was constrained to the categories of "responsive" and "non-responsive"; there was no category for "arguable" documents. Therefore, we cannot consider a post-hoc determination of "arguable" as necessarily contradicting the TA's original adjudication at TREC 2009.

Topic	Production Request
201	All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in structured commodity transactions known as "prepay transactions."
202	All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in transactions that the Company characterized as compliant with FAS 140 (or its predecessor FAS 125).
203	All documents or communications that describe, discuss, refer to, report on, or relate to whether the Company had met, or could, would, or might meet its financial forecasts, models, projections, or plans at any time after January 1, 1999.
204	All documents or communications that describe, discuss, refer to, report on, or relate to any intentions, plans, efforts, or activities involving the alteration, destruction, retention, lack of retention, deletion, or shredding of documents or other evidence, whether in hard-copy or electronic form.
205	All documents or communications that describe, discuss, refer to, report on, or relate to energy schedules and bids, including but not limited to, estimates, forecasts, descriptions, characterizations, analyses, evaluations, projections, plans, and reports on the volume(s) or geographic location(s) of energy loads.
206	All documents or communications that describe, discuss, refer to, report on, or relate to any discussion(s), communication(s), or contact(s) with financial analyst(s), or with the firm(s) that employ them, regarding (i) the Company's financial condition, (ii) analysts' coverage of the Company and/or its financial condition, (iii) analysts' rating of the Company's stock, or (iv) the impact of an analyst's coverage of the Company on the business relationship between the Company and the firm that employs the analyst.
207	All documents or communications that describe, discuss, refer to, report on, or relate to fantasy football, gambling on football, and related activities, including but not limited to, football teams, football players, football games, football statistics, and football performance.

Table 4: Mock production requests ("Topics") composed for the TREC 2009 Legal Track Interactive Task.

```
Date: Tuesday, January 22, 2002 11:31:39 GMT
Subject:

I'm in.  I'll be shredding 'till 11am so I should
haveplenty of time to make it.
```

Figure 1: A clearly responsive document to Topic 204. This document was coded as non-responsive by a contract attorney, although it clearly pertains to document shredding, as specified in the production request.

```
From:   Bass, Eric
Sent:   Thursday, January 17, 2002 11:19 AM
To:     Lenhart, Matthew
Subject:        FFL Dues

You owe $80 for fantasy football.  When can you pay?
```

Figure 2: A clearly responsive document to Topic 207. This document was coded as non-responsive by a contract attorney, although it clearly pertains to fantasy football, as specified in the production request.

of assessment guidelines.[7] We illustrate our post-hoc analysis using exemplar documents that were successfully appealed as responsive to topics 204 and 207. We chose these topics because they were the least technical and, therefore, the most accessible to readers lacking subject-matter expertise.

Figures 1 and 2 provide examples of documents that are clearly responsive to Topics 204 and 207, but were coded as non-responsive by the first-pass assessors. The first document concerns shredding, while the second concerns payment of a Fantasy Football[8] debt. We assert that the only reasonable assessment for both of these documents is "responsive."

Figures 3 and 4, on the other hand, illustrate documents for which the responsiveness to Topics 204 and 207, respectively, is arguable. Reasonable, informed assessors might disagree, or find it difficult to determine, whether or not these documents met the criteria spelled out in the production requests and assessment guidelines.

[7] The guidelines, along with the complaint, production requests, and exemplar documents, may be found at `http://plg1.cs.uwaterloo.ca/trec-assess/`.

[8] "Fantasy football an interactive, virtual competition in which people manage professional football players versus one another." `http://en.wikipedia.org/wiki/Fantasy_football_(American)`.

```
Subject: Original Guarantees
Just a followup note:
We are still unclear as to whether we should continue
to send original  incoming and outgoing guarantees to
Global Contracts (which is what we have  been doing
for about 4 years, since the Corp. Secretary kicked us
out of  using their vault on 48 for originals because
we had too many documents).  I  think it would be
good practice if Legal and Credit sent the originals
to  the same place, so we will be able to find them
when we want them.  So my  question to y'all is, do
you think we should send them to Global Contracts,  to
you, or directly the the 48th floor vault (if they let
us!).
```

Figure 3: A document of arguable responsiveness to Topic 204. This message concerns *where* to store particular documents, not specifically their destruction or retention. Reasonable, informed assessors might disagree as to its responsiveness, based on the TA's conception of relevance.

```
Subject:        RE: How good is Temptation Island 2
They have some cute guy lawyers this year-but I bet you
probably watch that manly Monday night Football.
```

Figure 4: A document of arguable responsiveness to Topic 207. This message mentions football whimsically and in passing, but does not reference a *specific* football team, player, or game. Reasonable, informed assessors might disagree about whether or not it is responsive according to the TA's conception of relevance.

7 Discussion

Our evidence supports the conclusion that responsiveness – at least as characterized by the production requests and assessment guidelines used at TREC 2009 – is fairly well defined, and that disagreements among assessors are largely attributable to human error. As a threshold matter, only 5% of the first-pass assessments were appealed. Since participating teams had the opportunity and incentive to appeal the assessments with which they disagreed, we may assume that, for the most part, they agreed with the first-pass assessments of the documents they chose not to appeal. That is, the first-pass assessments were on the order of 95% accurate. Second, we observe that 89% of the appeals were upheld, suggesting that they had, for the most part, a reasonable basis.

Our study considers only those appealed documents for which the appeals were upheld – about 89% of the appealed documents, or 4.5% of all assessed documents. Are these documents arguably on the borderline of responsiveness, as one might suspect? At the TREC 2009 Workshop, many participants, including the authors, voiced opinions to this effect. An earlier study by the authors preliminarily examined this question and found that, for two topics,[9] the majority of non-responsive assessments that were overturned were the result of human error, rather than questionable responsiveness [6]. The aim of the present study was to further test this hypothesis, by considering the other five topics, and also responsive assessments that were overturned (*i.e.,* adjudicated to be non-responsive). To our surprise, we found that we judged nearly 90% of the overturned documents to be clearly responsive, or clearly non-responsive, in agreement with the Topic Authority. We found another 5% or so of the documents to be clearly responsive or clearly non-responsive, contradicting the Topic Authority. *Only 5% did we find to be arguable*, indicating a borderline or questionable decision. Accordingly, we conclude that the vast majority of disagreements arise due to simple human error; error that can be identified by careful reconsideration of the documents using the production requests and assessment guidelines.

Our results also suggest that the TA assessments, while quite reliable, are not infallible. We confirmed this directly for Topic 204 by having the same TA reconsider ten documents that she had previously assessed as part of TREC 2009. For three of the ten documents, the TA contradicted her earlier assessment; for two of the ten, the TA coded the documents as arguable. For only half of the documents did the TA unequivocally reprise her previous assessment. While we did not have the TAs for the other topics reconsider their assessments, we are confident from our own analysis of the documents that some of their assessments were incorrect.

All in all, the total proportion of documents that are borderline, or for which the adjudication process yielded the wrong result, appears to be quite low. Five percent of the assessed documents were appealed; 90% of those appeals were upheld; and of those, perhaps 10% were borderline – that is, only about 0.45% of the assessed documents were "arguable." It stands to reason that there may be some borderline documents that our study did not consider. In particular, we did not consider documents that the first-pass assessor and the TREC 2009 participants agreed on, and which were therefore not appealed. We also did not consider documents that were appealed, but

[9] Topics 204 and 207, which were chosen because they were the least esoteric of the seven topics.

for which the TA upheld the first-pass assessment. We have little reason to believe that the number of such borderline documents would be large in either case; however, a more extensive study would be necessary to quantify this number. In any event, we are concerned here specifically with the *cause* of assessor disagreement that was observed, and since there is no assessor disagreement on these particular documents, this quantity has no bearing on the hypotheses we were testing.

We characterize our study as qualitative rather than quantitative for several reasons. The documents we examined were not randomly selected from the document collection; they were selected in several phases, each of which identified a disproportionate number of controversial documents:

1. The stratified sampling approach used by TREC 2009 to identify documents for the first-pass assessment emphasized documents for which the participating teams had submitted contradictory results;

2. The appeals process selected from these documents those for which the teams disagreed with the first-pass assessment;

3. For our post-hoc assessment, we considered only appealed documents for which the Topic Authority disagreed with the first-pass assessor; and

4. For our TA reconsideration, we considered only ten percent of the documents from our post-hoc assessment – those for which the post-hoc assessment disagreed with the decision rendered by the TA at TREC 2009.

All of these phases tended to focus on controversial documents, consistent with our purpose of determining whether disagreement arises due to ambiguity concerning responsiveness, or human error. Therefore, it would be inappropriate to use these results to estimate the error rate of either the first-pass assessor or the Topic Authority on the collection as a whole.

Finally, neither of the authors is at arm's length from the TREC 2009 effort; our characterization of responsiveness reflects our informed analysis and as such, is amenable to debate. Accordingly, we invite others in the research community to examine the documents themselves and to let us know their results. Towards this end, we have made publicly available the text rendering of the documents we reviewed for this study.[10]

8 Conclusion

It has been posited by some that it is impossible to derive accurate measures of recall and precision for the results of any document review process because large numbers of documents in the review set are "arguable," meaning that two informed, reasonable reviewers could disagree on whether the documents are responsive or not. The results of our study support the hypothesis that the vast majority of cases of disagreement are a product of human error rather than documents that fall in some "gray area" of responsiveness. Our results also show that while Topic Authorities – like all human

[10] *See* `http://plg1.cs.uwaterloo.ca/~gvcormac/maura1/`.

assessors – make coding errors, adjudication of cases of disagreement in coding using a senior attorney can nonetheless yield a reasonable gold standard that may be improved by systematic correction of the estimated TA error rate.

References

[1] BAILEY, P., CRASWELL, N., SOBOROFF, I., THOMAS, P., DE VRIES, A., AND YILMAZ, E. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (2008), ACM, pp. 667–674.

[2] BÜTTCHER, S., CLARKE, C., AND CORMACK, G. *Information retrieval: Implementing and evaluating search engines*. MIT Press, 2010.

[3] CHU, H. Factors affecting relevance judgment: a report from TREC Legal track. *Journal of Documentation 67*, 2 (2011), 264–278.

[4] CORMACK, G., AND MOJDEH, M. Machine learning for information retrieval: TREC 2009 web, relevance feedback and legal tracks. In *The eighteenth Text REtrieval Conference proceedings (TREC 2009), Gaithersburg, MD* (2009).

[5] EFTHIMIADIS, E., AND HOTCHKISS, M. Legal discovery: Does domain expertise matter? *Proceedings of the American Society for Information Science and Technology 45*, 1 (2008), 1–2.

[6] GROSSMAN, M. R., AND CORMACK, G. V. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology XVII*, 3 (2011).

[7] HEDIN, B., TOMLINSON, S., BARON, J. R., AND OARD, D. W. Overview of the TREC 2009 Legal Track. In *The Eighteenth Text REtrieval Conference (TREC 2009)* (2010). To appear.

[8] ROITBLAT, H. L., KERSHAW, A., AND OOT, P. Document categorization in legal electronic discovery: Computer classification vs. manual review. *Journal of the American Society for Information Science and Technology 61* (2010), 70–80.

[9] VOORHEES, E. M. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management 36*, 5 (2000), 697–716.

[10] WANG, J., AND SOERGEL, D. A user study of relevance judgments for E-Discovery. *Proceedings of the American Society for Information Science and Technology 47*, 1 (2010), 1–10.

Retrospective and Prospective Statistical Sampling in Legal Discovery[1]

Richard T. Oehrle, Cataphora Legal, a division of Cataphora, Inc.(rto@cataphora.com)

DESI IV Workshop

Statistical sampling can play an essential double role in defining document sets in response to legal discovery requests for production. Retrospectively, looking backward at existing results, statistical sampling provides a way to measure quantitatively the quality of a proposed production set. Prospectively, looking forward, statistical sampling (properly interpreted) shows how the quality of a proposed production set can be improved. The proposed improvements depend on transparency of classification: in order to correct clashes between human judgments of sampled data and a proposed hypothesis, one must know the source of the misclassification. And in correcting such misclassifications, one must take care to avoid standard dangers like overfitting.

Section 1 below sets the stage by presenting the most basic material on statistical sampling and introducing the concept of data profiles. Section 2 argues that statistical sampling in retrospective mode is the only practical way to assess production quality. Section 3 offers several reasons why statistical sampling assessment has failed to become the standard practice it deserves to be in the field of legal discovery. Section 4 focuses on the use of statistical sampling in prospective, forward-looking mode, to drive iterative improvement. Section 5 describes at a high level some of our practical experience at Cataphora using iterative statistical sampling to force rapid convergence of an evolving responsiveness hypothesis with very high quality standards of review assessment. (In fact, the intrinsic role that statistical sampling plays in this process described in this section—a process that has consistently yielded measurably high quality—is one reason to consider the issues that arise in the preceding sections.) Along the way, we offer a variety of questions for discussion in a series of footnotes.

1 Background

1.1 statistical sampling, confidence intervals, confidence levels

Statistical sampling starts with a sample drawn from a data set. We cannot be certain that the sample is representative of the data as a whole. But we can estimate the likelihood that it is. This estimate takes the form of two hedges—a confidence interval and a confidence level. The confidence interval pads the particular results derived from the sample with room for error on both sides (say +/- 5%). The confidence level states how probable it is that any sample drawn from the data will fit within this interval. Intuitively, think of any distribution as being roughly like a bell curve, which prototypically has a central axis (the vertical line that goes through the top of the bell), with distribution falling off symmetrically on either side. Then think of a sample as a subset of the region between the horizontal x-axis and the bell curve. If the sample is randomly selected, because of the shape of the bell curve, most of the items in the sample fall within a relatively small interval flanking the central axis symmetrically on either side. This interval is represented by the confidence interval, and when the bell curve is relatively normal—not too flat—most of the points are not far from the central axis. Most is not the same as all, of course. And the confidence level is added to

[1]I'd like to thank the anonymous DESI IV referees for their constructive comments. Of course, any errors in this paper are my responsibility, not theirs.

deal with the outliers on either side that don't make it into the interval. (These outliers form the tails on either side of the bell that trail off on either side getting closer and closer to the x-axis the further away they are from the central axis.) If we claim a 95% confidence level, the claim is roughly that at least 95% of the points under the bell curve fall within the window and less then 5% of the points under the bell curve fall within the outlying tail on either side outside the window. This is why we need both a confidence interval and a confidence level.

1.2 data profiles

Many discussions of information retrieval and such basic concepts as *recall* and *precision* assume a binary distinction between *responsive* (or *relevant*) and *non-responsive* (or *non-relevant*). As anyone with any practical experience in this area knows, this is quite an idealization. To get a grasp on the range of possibilities that emerges from combining a responsiveness criterion with a dataset, it is useful to introduce the concept of a *data profile*.[2] Suppose we are given a dataset D and some omniscient being or oracle has the quick-wittedness and charity to rank every document on a scale from 0 (the least responsive a document could possibly be) to 10 (the most responsive a document could possibly be), and to provide us with a list of documents ranked so that no document is followed by a document with a higher rank. Here are some illustrative pictures, where documents are represented as points on the x-axis (with document d1 corresponding to a point to the left of the point corresponding to document d2 if document d1 precedes document d2 in the oracle's list), and the degree of responsiveness of a document represented by points on the y axis from 0 (least responsive) to 10 (most responsive).

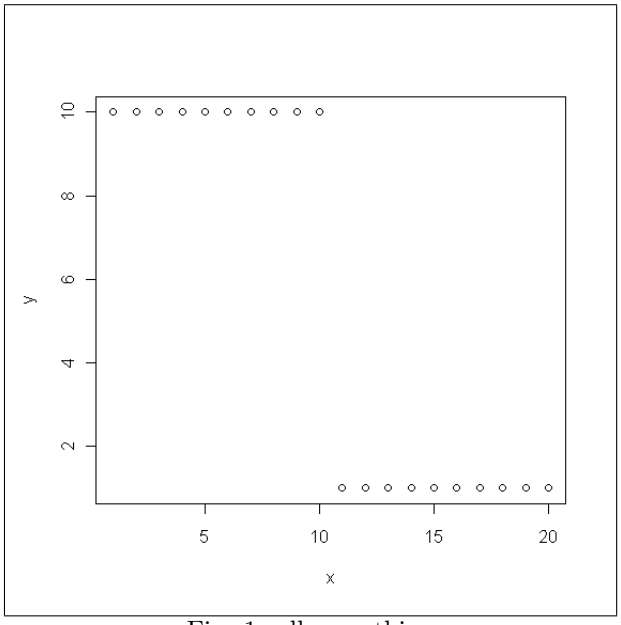


Fig. 1: all-or-nothing

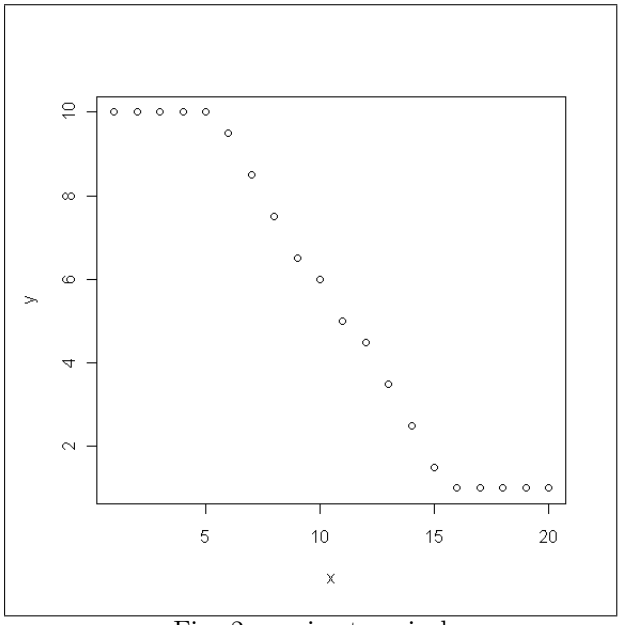


Fig. 2: semi-categorical

[2]The intuitions behind this concept have affinities with the AUC Game described by Gordon Cormack & Maura Grossman (2010).

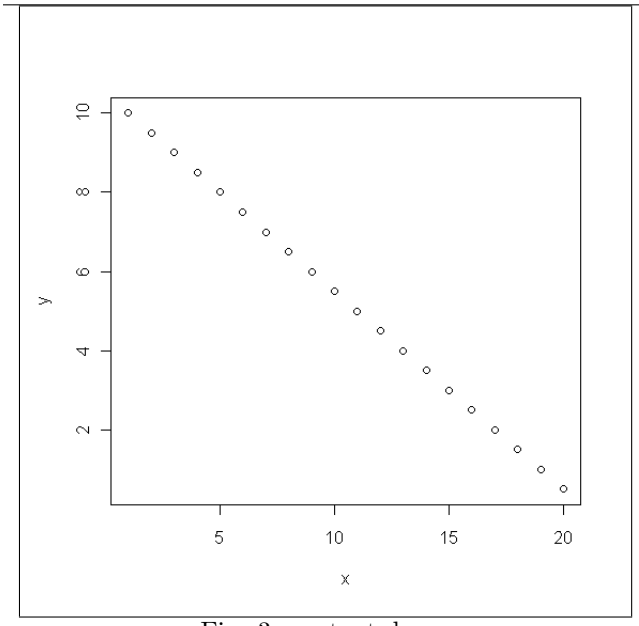


Fig. 3:constant decay

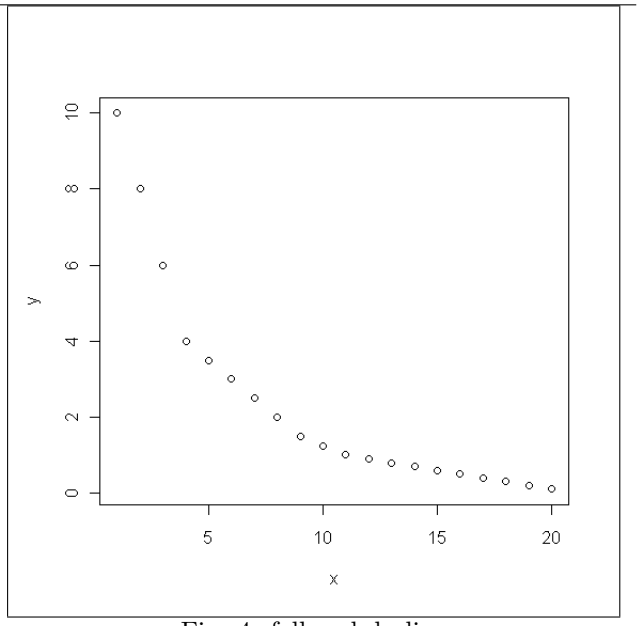


Fig. 4: fall-and-decline

Technically, the fundamental property of these graphical displays is that the relations they depict are weakly decreasing: if a point (x_1, y_1) is to the left of a point (x_2, y_2) (so that $x_1 < x_2$)), then y_2 cannot be greater than y_1. The point of introducing them is simple: they make it possible to apprehend and explore a landscape of theoretical possibilities which illuminates the practical questions that practitioners face.[3]

2 Statistical Sampling is the only practical way to assess production quality

Legal Discovery requires a specification (at some level of detail) of what is regarded as Responsive and what is regarded as Non-Responsive with respect to a particular document collection. A production set or potential production set drawn from this collection can be regarded as a hypothesis about which documents satisfy the Responsive specification.

When both the underlying dataset and the proposed production set are relatively large,[4] there is only one practical way to assess the quality of such an hypothesis: statistical sampling. Statistical sampling relies on a solid and well-understood mathematical foundation. It has been employed extensively across a broad range of subject matters. It is quantitative, amazingly efficient, replicable, and informative.

[3] **Question:** Given a data profile associated with a responsive criterion R associated with a request for production and a dataset D, what portion of the dataset should be produced?

[4] A referee has noted the potential and importance of hot-document searches, whose relative rarity may insulate them from the representative sweep of statistical sampling. We will come back to this point briefly in the final section.

2.1 human categorization does not in and of itself entail quality results

It is sometimes assumed that a quantitative assessment of production quality is unnecessary, on the grounds that the method used to define the candidate production entails its high quality. But assessment is completely independent of this process of definition. If we define a candidate production set by flipping a fair coin, we should still be able to assess the quality of the result. Historically, human manual review of the entire document collection has served as a benchmark of sorts, based on the assumption that human manual review must be correct. But there have always been skeptics who have doubted the efficacy of human manual review. And empirical investigations, which are not easy to arrange in practice, are beginning to show this assumption is in fact incorrect: defining a potential production set by human manual review does not guarantee a high quality result. (See, for example, Roitblat, Kershaw, and Oot (2010).)

Recently, there has been another version of this argument applied to automated methods of review. This version takes a form like the following: if a method can be shown to be consistently accurate across a diverse population of document collections, then we can assume that it will be consistently accurate when applied to a new collection that it has never been tested on. This formulation involves some delicate conditions concerning the properties of the document collections that form the provisional testing set. How could one be sure in practice that these same conditions actually hold when we move to a new document collection? The simplest way is to measure the quality of the results by statistical sampling. But in this case, it isn't necessary to rely on the delicate conditions at all: the quantitative assessment will provide the information needed.

2.2 human re-review: expensive, inefficient

One conceivable way to test quality is to re-review the entire document collection manually. This approach would be expensive and time-consuming. Furthermore, recent empirical research (such as Roitblat, Kershaw, and Oot (2010)) shows that multiple human reviews of the same document set yield astonishingly large disagreements in judgments.[5] In other words, apart from its expense and inefficiency, this approach is unlikely to provide a true assessment of the quality of a proposed production set.

Similarly, suppose the candidate production set was defined by a fully automated process. We can't test the process by re-running the fully automated process. If the process is consistent, then a second run will replicate the results of the earlier run, without providing any information about quality. Again, the remedy is to submit the results to statistical sampling.

2.3 informal QC vs. statistical sampling

Statistical sampling is sometimes replaced by informal browsing through a candidate production set. This differs from the statistical approach in a number of ways. For example, the sample set is not always selected appropriately. Moreover, quantitative results are not always tabulated. While the results of this seat-of-the-pants QC can be better than nothing, they do not provide the detailed insights available from statistical sampling.

[5] **Question:** if we consider a data profile associated with the responsive criterion and data set of the Verizon study, what is the corresponding *error profile*: that is, are clashes in judgment randomly distributed across the x-axis values? are they concentrated at the extremes of responsiveness / nonresponsiveness (represented by the left end and right end, respectively? are they concentrated in the middle? ...

2.4 if human review is fallible in general, why is it effective in sampling?

There are two practical reasons to distinguish the general properties of human review in large linear reviews and the general properties of human review in sampling reviews. First, because sampling review is remarkably efficient, it makes sense to employ senior attorneys with knowledge of both the details of the case at hand and the underlying law, rather than junior associates or contract attorneys or paralegals. (Compare the role of the *Topic Authority* in recent TREC Legal rounds.) In other words, the population is different, in a way that should (in principle) tilt the balance toward improved results. Second, our knowledge of the inconsistencies of multiple human reviews is based on large datasets with thousands of judgments. Since sampling reviews involve a much smaller dataset, clashes between a given reasonable hypothesis concerning responsiveness and actual expert reviewer judgments tend in practice to be even smaller. In fact, they are small enough to be subjected to individual examination, which sometimes confirms the expert reviewer, but at other times confirms the given hypothesis. This kind of detailed examination provides a highly valuable constraint on the quality of information provided by human reviewers, a constraint absent in the large scale multiple reviews whose differences have been studied. Finally, sampling review occurs over a time-span that lessens the risks of fatigue and other vicissitudes.

2.5 summary

In summary, if you want to know how good your proposed production set is, statistical sampling provides a quantitative, replicable, efficient, informative, practical, defensible answer. No other method known to us comes even close.[6]

3 Why isn't statistical sampling the de facto standard in legal discovery?

Properly conducted statistical sampling answers basic questions about the quality of legal discovery productions (up to approximations represented by confidence interval and confidence level). Why doesn't statistical sampling play a more central role when issues concerning discovery arise? Why isn't it regarded as reasonable and customary?

3.1 is no quantitative check needed?

One possible answer to this question (already discussed above) is that no quantitative check on quality is needed and if it isn't needed, it poses an additional and unnecessary burden. The primary justification for this answer is that the particular method chosen (manual, technology assisted, or fully automatic) serves as a guarantee of production quality. But empirical studies of manual review consistently show that manual review does not support this justification. And there is little reason to think that automated forms of review fare better. Moral: skepticism is called for.

[6] **Question:** where in the E-Discovery process should statistical sampling be employed? Example: if one side proposes to use *keyword culling* to reduce the size of the data and the associated costs of discovery, should the other side be provided with quantitative measures of the impact of this procedure on the responsive and non-responsive populations before and after culling?

3.2 what is a practical standard?

A related myth is that on the assumption that human review is perfect (100% recall and 100% precision), revealing actual sampling results will introduce quantitative figures that can never meet this perfect standard. It's true that statistical results always introduce intervals and confidence levels. And while such results can approach 100%, sampling can never guarantee 100% effectiveness. The practical impact of these facts is that some may feel that introducing statistical sampling results can only serve to illuminate defects of production. But if the introduction of statistical sampling results were the accepted practice, whether for manual or automated forms of review and production, it would very quickly become clear what the acceptable numbers for review quality actually are, what numbers require additional work, and what numbers are of high enough quality that further improvements would require increasing amounts of work for decreasing rewards.[7]

3.3 ignorance may be preferable to the consequences of knowledge

There is another possible factor which may have contributed to the failure of statistical sampling to be regarded as a reasonable and customary part of discovery. This factor has nothing to do with disclosing such results to the court or to other parties. Rather, it involves fear that the results will not satisfy one's own standards. Weighed in the balance, fear and ignorance trump knowledge and its consequences. Suppose you conduct a traditional linear review on a large document set. At the end of the review, you sample appropriately across the dataset as a whole to estimate the recall and precision of your candidate production. What if you were aiming for 90% at a minimum (with a confidence interval of 5% and a confidence level of 95%), but your sampling review shows that the recall is 75%. What choices do you face? Do you certify in some way a review that is plainly deficient in its results (even though it may have been conducted flawlessly)? Do you launch the manual review again from scratch, with all the attendant costs in time, effort, and money—and no guarantee in advance that the results of the second round of review will outperform the unsatisfactory results of the first review? One way to avoid this dilemma is to refrain from a quantitative estimate of production quality. The resulting shroud of ignorance obscures the painful choice.

This situation is not restricted to cases involving human manual review. For example, a similar dilemma would arise in circumstances in which the initial results depended on a black-box algorithm—that is, an automated approach that offers a hypothesis about how documents are to be sorted in a Responsive set and a Non-Responsive set, but does not reveal the details of how the hypothesis treats individual documents. For example, think of clustering algorithms that can be adjusted to bring back smaller or larger clusters (by strengthening or relaxing the similarity parameters). In the face of unsatisfactory recall results, one might be able to adjust the algorithm to drive recall numbers. Typically, however, this very adjustment adversely affects precision numbers, because additional documents that are in reality non-responsive may be classified as Responsive.

[7] **Question:** who should have access to sampling numbers? Example: does the counsel for the sampling side want to know that the sampling numbers are not perfect—they never are, for reasons discussed above—even they may far exceed contemporary standards? **Question:** what role do sampling measures play in defending production quality before a judge? Example: can the opposing side reasonably demand statistical measures of recall and precision when the quality of a production to them is in question?

3.4 making reality your friend

Not every method of defining a production set faces this dilemma. In the next section, we discuss the conditions needed to leverage statistical sampling results to improve review quality. And subsequently, because the necessary conditions are somewhat abstract, we discuss our experience at Cataphora using this iterative review model over the past seven years.

4 Leveraging statistical sampling results prospectively for hypothesis improvement

If the results of statistical sampling can be used to improve a hypothesis about a potential production set—that is, improve recall and improve precision—then a system based on successive rounds of sampling and hypothesis refinement can return better and and better results.[8] Before considering how quickly this convergence takes place in practice in the next section, we focus first on exactly how sampling can be leveraged for hypothesis improvement.

Suppose you review 800 documents selected to test the recall of your current hypothesis. This is a test of completeness, whose goal is to ascertain whether the current hypothesis mischaracterizes Responsive documents as Non-Responsive. Suppose that the resulting review judgments are as follows:

	hypothesized Responsive	*hypothesized NonResponsive*
judged Resp	80	40
judged NR	40	640

This sampling review thus confirms that the current hypothesis is underperforming with respect to recall: 40 documents that were hypothesized to be NonResponsive were judged Responsive. We might think that 40 is a relatively small number: only 5% of the 800 document sample. But this represents a third of all the documents judged Responsive in the sample. Suppose that the document set as a whole contains 800,000 items. If the sample is representative, 120,000 of them are responsive and our current hypothesis only identifies 80,000 of them. This is clearly unacceptable. The hypothesis needs to be revised and substantially improved.

Let's assume that all the clashes between review judgments and the current categorization hypothesis are settled in favor of the review. (In practice, it happens that further scrutiny of clashes can lead to a resolution that favors the categorization hypothesis rather than the human review.) Having determined the identity of 40 false negatives, the least we can do is to ensure that these 40 are re-categorized in some way so that they are categorized as Responsive. But this is obviously insufficient: the 40 false negatives are representative of a much larger class. It's this larger class that we must be concerned with and we want to use information extractable from the 40 false negatives to improve our hypothesis. What we seek is a way of categorizing these 40 false negatives that generalizes appropriately over the data as a whole. Two basic cases arise.

In the first case, the 40 false negatives are categorized by the current hypothesis, but the combination of the categorization components involved are incorrectly associated with NonResponsiveness, rather than Responsiveness. By adjusting the way in which such combinations of categorization

[8] See the work by Grossman & Cormack, cited earlier, and Büttcher, Clarke, and Cormack (2010), especially section §8.6 *Relevance Feedback*. What we describe here is a form of iterated supervised feedback involving both relevant and non-relevant information.

components determine Responsiveness or NonResponsiveness, the performance of the current categorization hypothesis can be improved in a suitably general way. (This is the very situation that Cataphora's patented Query Fitting Tool is designed to address, particularly when the number of categorization components is so high that finding a near-optimal combination of them manually is challenging.)

In the second case, the 40 false negatives are not categorized at all or are categorized in a way that is overly specific and not suitable for generalization. In this case, we seek to divide the 40 documents into a number of groups whose members are related by categorization properties (topics, subject line information, actors, time, document type, etc.). We next add categorization components sensitive to these properties (and independent of documents known to be NonResponsive) and assign documents satisfying them to the Responsive class.

The result of these two cases is a revised categorization hypothesis. It can be tested and tuned informally during the course of development. But to determine how well performance has improved, it is useful to review an additional sample drawn in the same way. If the results confirm that the revised categorization hypothesis is acceptable, the phase of hypothesis improvement (for recall, at least) can be regarded as closed. (When all such phases are closed, a final validation round of sampling is useful to ensure overall quality.) On the other hand, if the results suggest that further improvements are indicated, we repeat the tuning of the categorization hypothesis as just outline and test a subsequent sample. In this way, we get closer and closer to an ideal categorization hypothesis. In practice, we get a lot closer with each round of improvement. (Details in the next section.)

4.1 transparency

There is one critical point to note about this iterative process: it depends critically on the transparency of categorization. In order to improve a categorization hypothesis, we need to know how particular documents are categorized on the current hypothesis and we need to know how these documents will be categorized on a revised hypothesis. If we cannot trace the causal chain from categorization hypothesis to the categorization of particular documents, we cannot use information about the review of particular documents to institute revisions to the categorization hypothesis that will improve results not only for the particular documents in question but also for more general sets of documents containing them.

5 Cataphora's practical experience: empirical observation on the success of iterative sampling

We've shown above how statistical sampling results can be used to both measure performance and drive hypothesis improvement. If we follow such a strategy, results should improve with each iteration. But how much better? And how many iterations are required to reach high-quality results? In this section, we address these questions from a practical, empirically-oriented perspective. Cataphora Legal has been successfully using statistical sampling to measure performance and to improve it for almost a decade. In what follows, we draw on this experience, in a high-level way (since the quantitative details involve proprietary information). Our goal is not to advertise Cataphora's methods or results, but to document the effectiveness of statistical sampling review in the development of high-quality hypotheses for responsiveness categorization.

Before discussing project details, it is worth pointing out that statistical sampling can be integrated with other forms of review in many ways. As an example, it may be desirable and prudent to review the intersection of a responsive set of documents and a set of potentially privileged documents manually, because of the legal importance of the surrounding issues. As an other example, it may be desirable to isolate a subpopulation of the dataset as a whole to concentrate for manual hot-document searches. In other words, different techniques are often appropriate to different subpopulations. Overall, such mixed methods are perfectly compatible.

In a recent project, we developed a hypothesis concerning responsiveness for a document set of over 3 million items. This work took approximately 2 person-weeks, spread out over 3 months (not related to our internal time-table). Attorneys from the external counsel reviewed a randomly selected sample of the dataset. The recall of our hypothesis exceeded 95%. Precision exceeded 70%.

Not long after this sampling review occurred, we received additional data, from different custodians, as well as some modifications in the responsiveness specification. After processing the new data, we arranged a sampling review, using the previously developed responsiveness hypothesis. In this second sample, involving significant changes to the population, the performance of the original responsiveness hypothesis declined considerably: recall dropped to about 50% from the previous high on the original data. We spent days, not weeks, revising and expanding the responsiveness hypothesis in ways that generalized the responsive review judgments in the sampling results. At the end of this process, the attorneys reviewed a fresh sample. Results: the recall of the revised hypothesis exceeded 93%; precision exceeded 79%.

These numbers compare favorably with publicly available estimates of human manual review performance. The dataset involved was large. The overall process was efficient. In the present context, what is most notable is that the convergence on high quality results was extremely quick and the role played in this convergence by statistical sampling was significant.

References

Büttcher, Stefan, Charles L. A. Clarke, and Gordon V. Cormack, *Information Retrieval: Implementing and Evaluating Search Engines*, Cambridge: The MIT Press, 2010.

Gordon V. Cormack and Maura R. Grossman, *TREC Legal Track—Learning Task Draft Guidelines*, `http://plg.uwaterloo.ca/~gvcormac/legal10/legal10.pdf`, (2010).

Roitblat, H., A. Kershaw, and P. Oot, 'Document categorization in legal electronic discovery: computer classification vs. manual review', *Journal of the American Society for Information Science and Technology*, 61.1, pp. 70-80, 2010.

An Intelligent Approach to E-discovery

Steve Akers

CTO/Founder Digital Reef Inc.

Boxborough, Ma

Jennifer Keadle Mason, Esq.

Partner Mintzer, Sarowitz, Zeris, Ledva & Meyers, LLP

Pittsburgh, Pa

Peter L. Mansmann

CEO Precise Litigation Inc.

Pittsburgh, PA

Introduction

Legal Discovery and assessment is an expensive proposition for corporations and organizations of all types. Last year (2010) it is estimated that $1 billion – $3 billion was spent on legal discovery processing alone[1.] This cost is large and growing; finding more intelligent methods to assess Electronically Stored Information (ESI) and understand what is contained within it is a goal of not just corporate personnel but also lawyers and legal service providers (companies providing legal discovery services). This paper outlines a proposed "standard" methodology and defines "ideal" tools and technology methods that combined are suggested as a "standard" for search. It focuses on 1. a standard methodology to identify potentially relevant data, and 2. tools and technology that can aid this process. It discusses important technological aspects of Ediscovery and how products either address or fall short of perfection in certain areas. Using the best process of identification in combination with the proper technologies for specific data types in order to have resulting cost-effective Ediscovery is the focus of this paper.

One of the quandaries facing attorneys is how best to approach any particular data set to identify potentially relevant information either for their own use or use in responding to discovery. Increasing variety of data types and sources along with expanding volumes of unstructured data has made the decision of how to search the data more imperative than ever. Analytic search tools have blossomed in this environment and certainly provide some of the best options for searching. However analytic searching has many flavors in and of itself. Understanding the pros and cons to each approach is important in deciding which route to go. In addition attorneys cannot ignore "traditional" search methods as they can be an effective supplement to analytic searching or in some cases may be the best primary method for running a search. The decisions about which route to take is largely driven by the types of data being searched, the relative organization of the data being searched, the particularity of the case facts, and the attorneys familiarity with the case facts and client.

The application of keywords has long been the standard for searching data sets. Keyword searching in its basic form is identifying any documents that contain particular terms. Ideally the parties discuss the keywords to be run, review a report of the initial search results to discuss any necessary adjustments, apply a privilege filter, review, and produce. These steps may be repeated numerous times to allow the parties to apply new search terms based upon the knowledge gained in reviewing the records. The problems with keyword searching are several and include: The parties must have sufficient knowledge of the case facts and industry/party parlance; straight keyword searching will not find misspellings; natural language usage has the problem of synonymy (multiple words with the same meaning – kitten, cat, feline) and polysemy (same word having different meanings – strike); finding variations of people's names can be difficult (Dr. Jones, Indiana Jones, Indiana J.).

Because of these difficulties in running straight keyword searches, variants on the searching were developed to work around some of the deficiencies. Attorneys began running keyword searches in conjunction with metadata searches. Star searching allows the user to find root words to account for variations (interp* - would find interpret & interpretation). Fuzzy searching allowed users to find words within a certain percentage similarity of the word being searched. Proximity searching allowed

[1] Source: <Marketing to supply reference to report by Gartner or Forrester>

users to search for words within a certain distance of other words of each other. These variants on the keyword search alleviated some of the issues discussed above, but still didn't overcome the obstacles of synonymy and polysemy. This is where analytic searching has come to the forefront.

Analytic searching in, its most rudimentary explanation, is a method of finding or grouping documents based upon the content of the documents themselves not solely on a keyword(s) being used. This is commonly employed by internet search engines that allow the user to type in a basic subject inquiry and retrieve search results that aren't solely driven by the words entered into the search box. The basis for this search technology is the conversion of a document's contents into numeric values that allows the computer to compare differing document's values in order to determine similarity of content. By approaching document comparison in this way, specific terms (or even language) of a record becomes irrelevant to the determination of similarity.

Alex Thomo an Associate Professor in the Department of Computer Science at the University of Victoria offers the following example to explain the basis for how analytic searching (and in particular a Latent Semantic Analysis) operates in determining documents responsive to a search request:

> Suppose there is a set of five documents containing the following language:
>
> > Document 1: "Romeo and Juliet"
> >
> > Document 2: "Juliet: O happy dagger!"
> >
> > Document 3: "Romeo died by dagger."
> >
> > Document 4: "Live free or die - New Hampshire's motto"
> >
> > Document 5: "Did you know, New Hampshire is in New England?"
>
> A search is conducted for: ***dies, dagger***
>
> A classical IR system *(for our purposes keyword searching)* would rank d3 to be the top of the list since it contains both *dies*, *dagger*. Then, d2 and d4 would follow, each containing a word of the query.
>
> However, what about d1 and d5? Should they be returned as possibly interesting results to this query? A classical IR system will not return them at all. However (as humans) we know that d1 is quite related to the query. On the other hand, d5 is not so much related to the query. Thus, we would like d1 but not d5, or differently said, we want d1 to be ranked higher than d5.
>
> The question is: Can the machine deduce this? The answer is yes, LSA does exactly that. In this example, LSA will be able to see that term *dagger* is related to d1 because it occurs together with the d1's terms Romeo and Juliet, in d2 and d3, respectively.
>
> Also, term *dies* is related to d1 and d5 because it occurs together with the d1's term Romeo and

d5's term New-Hampshire in d3 and d4, respectively.

LSA will also weigh properly the discovered connections; d1 more is related to the query than d5 since d1 is "doubly" connected to *dagger* through Romeo and Juliet, and also connected to *die* through Romeo, whereas d5 has only a single connection to the query through New-Hampshire.

Using the above as an example, its apparent that analytic search engines can have a significant role in searching by obviating some of the problems of straight keyword searching. However, this does not mean that analytic searching alone will always be the most defensible method of searching. In addition when running analytic searching, its important to understand the different analytic engines and the limitations of each.

With all that said, in an attempt to identify a standard search process, this paper will first identify the problem, that is to determine what data you have and who has it. Next, the paper will elicit standard characteristics of a proposed standard search process. These will include identification methods and search and process methodologies to identify potentially relevant data in an effective, efficient and repeatable manner. Finally, the paper will discuss why those search and process methodologies are suggested for the search standardization model proffered. (The reasons for the identification methods proffered have been discussed in many articles, blogs and cases. Therefore, they are not discussed herein.)

Deciding What You Have (where you have it and how much)

The first problem with Ediscovery projects is assessing the magnitude and characteristics of the data in common knowledge repositories (email archives, SharePoint repositories, etc.). IT or Litigation Support professionals know they have a lot of data but are not sure where they have it and what these repositories contain. Not understanding the locations of data in an organization may seem like an odd statement but, for example, departments put SharePoint servers into production and users copy data to shared drives on networked file systems without knowing what they have copied. In other circumstances, administrators may not have knowledge of what systems are used for what data. These types of problems are ubiquitous and growing. The first step to effective assessment of a potential legal matter is to know what exists in various repositories within an organization. This is often the biggest problem in Ediscovery; identifying how much data exists and where it exists.

Legal Problem: Identification of Potentially Relevant Data

When litigation is filed and/or is reasonably anticipated, parties and counsel are required to identify and preserve data that is potentially relevant to the claims or defenses of the parties. In order to meet this obligation, parties have utilized numerous methodologies with varying levels of success. Success is often dependant upon the participants knowledge of the location of data as well as their understanding of the legal requirements and the technology involved. However, there has been no standard method to accomplish the task of searching for and reliably and efficiently locating that data.

Deciding who has it

Another big problem is in knowing who owns (or is responsible for) the information that is stored in various repositories. When a custodian (or potential custodian) is identified, it is important to know where their data might reside. Many organizations don't have any idea who owns what data and how often the data is accessed (if ever).

Technological Problem: Lack of Ownership insight

Historically data indexing and search solutions have not provided support for just a quick scan of file ownership in a short period of time to show what data requires further deeper analysis. Historically data has required full content indexing and analysis to provide insight into what it contains. Often the first level of analysis should be just a "who owns what" look at available data. In this case not all content needs full indexing. An intelligent approach is to perform a first-level analysis with just Meta data indexing and to then identify what content needs full content indexing. These are different "representations" of the data; one in Meta data form and one with all the content in the documents represented within the index. Systems with the ability to "represent" data in various ways let (users) reviewers decide what to deeply analyze. This saves time, storage space and lots of money.

Deciding What to Look For

A legal complaint will contain key facts about the case that will get the lawyers started on what they should ask the legal counsel representing an organization to identify and produce. A set of analytics that can assess the main complaint language or other "known key terms" and use these data to help build a set of "similar" documents would be very valuable to legal staff working on a case. Analytic processes that can expose "terms of interest" within a document to help lawyers involved with the case decide what to look for in other documents would be of great assistance to legal reviewers. Analytics to identify content that is "similar" to known example content is also very valuable.

Legal Problem: Identification of Potentially Relevant Claims/Defenses/Data

Upon identification of potential litigation and/or receipt of an action that has been filed, counsel must identify potentially relevant data and preserve it. How to most efficiently and effectively accomplish this goal is the problem faced by counsel and vendors alike. The proposed standard search methodology would begin with a litigation hold which involves identification of the "relevant topics" for the case. Relevant topics include but are not limited to the claims or defenses. Relevant topics might also include particular areas of interest for the litigation including, for example, profits/losses, prior claims, prior knowledge, investigations, testing, etc. in products liability cases. Once the relevant topics are known, the next area of inquiry is to identify the key players who might have possession of and/or who have created potentially relevant data. Key player questionnaires should be sent to these individuals. The questionnaire seeks information from the key player about "basic" data of which they are aware, why they were named, what position they hold, time frames of relevance, what documents they create in that position that might be relevant to the known relevant topics and where they store that data. It also should contain basic questions about the media on which they store information and where it is mapped and/or backed up. After this information is identified, a data map for the litigation should be drafted and key player interviews held. The interviews are usually more productive in person

where information sources located in the office, but often forgotten, can be identified. The interviews should be a much more detailed analysis of the way the corporation works, where data is stored, to whom it is copied, the purpose for which it is created, etc. The locations of the known potentially relevant data should also be discussed and, if possible, the path followed to locate specific server, drive, file names, etc. The data map for the litigation should be updated with this information and the client should verify the information contained therein by signature. Once the specific known locations are identified and all relevant topics have been discussed, known relevant documents can be pulled for use in creating better search parameters for further collection of data. In addition, once additional known relevant documents are located through the analytical search processes, the information from those documents can be utilized to search for other potentially relevant documents. Further, the terms/phrases from these new documents can be compared to the search results, i.e. clustering, to more efficiently identify potentially relevant data. In other words, the process should be iterative. In the meantime, an IT key player questionnaire should be sent to the person responsible for IT to determine the data architecture of the entity and backup/legacy information. The identification of mapping should also be sought along with information as to third party entities who maintain data and/or website information. Finally, IT and all key players should be asked to discontinue any document destruction, turn off auto delete and auto archive, and identify backup rotation. Potentially relevant data should be properly preserved depending upon data type and business capability until further decisions are made.

Technological Problem: lack of an "analytics toolkit" and Lack of Flexibility and Scale

Vendors have historically pushed one approach or solution on customers for Ediscovery. Every solution requires a search capability; when the solutions begin to contain analytics the vendor approach has been to offer a single type of analysis. One type of analysis does not always product the best results with all data sets. Sometimes email is the only source of data pertinent to a matter. One set of tools for email analysis may work fine for such a case. With other data pertinent to the same case, key evidence may exist in MS Word documents and email analysis techniques are not appropriate. This fact of life in Ediscovery has caused legal reviewers to turn to multiple solutions that are stand-alone applications. Moving data into and out of these applications introduces complexity and potential for error (human and otherwise). One platform providing a number of analytic tools that are appropriate at various times throughout the lifecycle of a case would be the most efficient approach to take for legal discovery. In addition, historically data indexing and search solutions lack the flexibility and scale to analyze the amount of data that may exist within a typical organization. A platform that could analyze large volumes of data efficiently would be helpful.

Deciding who shared what (and with whom)

Conversational analytics are very important to an Ediscovery solution. Knowing who spoke with whom about certain topics is often the cornerstone to legal analysis.

Technological Problem: lack of capability or full-featured capability for conversations

Some solutions use email header analysis, others use Meta data analysis and header analysis, others rely on message content. A solution that can identify content and header similarity is often the best solution. Providing this capability at scale is a challenge in many solutions.

Solution: A "Standard" Ediscovery Process

The solution to many of these problems with Ediscovery would be contained within an "standard e-discovery system" that connects to many sources of local data (behind the corporate firewall), to help litigation support personnel generate reports about data that may prove relevant to a case matter. This software would also interface with collection tools for desktop or laptop collection and process data at great scale in large data center environments. The ideal discovery system would also perform a number of functions that would allow collection processing and analysis of data regardless of file format. The system would support a system of "describing" data without moving it into a separate repository; reducing the required storage space to use the system and making collection efforts more targeted and specific; let alone more cost effective (take just what you need for legal hold for example). This "system" would be coupled with additional standard processes and best practices to form the "standard" Ediscovery process.

Such a system would also provide specific access to certain data items but not others based on user credentials and group membership of users (multi-tenancy; or the ability of multiple groups to use the system but only see specific documents depending on their role in the organization or on the review team). Please see Figure One and Figure Two (below) for an illustration of these concepts and how the system is deployed within an organization.

At the present time, it is not believed that any one platform on the market has all of the capabilities mentioned herein and certainly does not account for capabilities not yet developed. Counsel should always keep abreast of technological advances and incorporate the same into any standard process. Depending upon the case and the data set, you may want to consider one or more platforms that best fit your needs. The choice of platform may be driven by which of the following options, beyond standard keyword-Boolean options, are available and/or needed for your data set:

1. Platform capability to allow unprecedented scale of indexing, search and analytics
    a. OCR conversion to text capabilities to ensure that content is captured even in image files
    b. Exception processing of certain file types that may need special processing like forensic video or audio analysis
    c. Processing with "flexible attribute selection"
        i. Indexing with Regular Expression matching turned "on" or "off"
        ii. Numerical content turned "on" or "off"
2. Multiple representations of data corpora
    a. File system- level Meta data only
    b. Application-level Meta data
    c. Full content
    d. Analytic attribute structures for semantic analysis
    e. Analytic Meta data structures for user-supplied attributes "tagging"
    f. Analytic Meta data structures for machine generated attributes
        i. Cluster associations for similar documents

ii. Near-duplicate associations for similar documents
iii. Group views for search associations

3. Analysis Capabilities
    a. To help identify keyword criteria – figure out which words are contained within the data universe and subsequently determine which are most relevant
    b. To identify relationships in the content that is in need of scrutiny or discovery (clustering)
    c. To organize documents and relate keyword searches to content that is in an analytic folder
    d. To remove duplicate content from responsive document sets
    e. To identify versions of content within sets of documents (versions of contracts or emails)
    f. To identify language characteristics of documents (language identification)
    g. To identify email conversations and conversation "groups"
    h. Linguistic analysis (categorization in terms of meaning)
    i. Sampling to pull data from known locations to use for additional searching
    j. Supervised classification or categorization (using known relevant documents to form search queries to find other potentially relevant documents
    k. Lexical analysis (entity extraction or analysis)
4. Validation Capabilities (Whether in the platform or extraneous)
    a. To validate the search (pulling random sample of all documents to validate search methodology
    b. To validate the review for:
        i. Privilege
        ii. Confidential Information (i.e. other products, social security numbers)
        iii. Tagged/relevant topics (pulling random sample of reviewed data to validate the review process)

Definitions of Key Terms

Key terms relevant to understanding an ideal Ediscovery system are:

Representation of Data

In the ideal system, it is important to represent documents so that they can be identified, retrieved, analyzed and produced for attorney review. Documents can be represented within the system by some sort of index or by certain kinds of data structures (covered in detail in a later section of this document). Different types of analysis require different types of indices or data structures. It is ideal to build the appropriate data structures to support the kind of data analysis that is required at a certain stage of the ediscovery process.

In an ideal system document representations can be constructed to include certain kinds of information but not other types. This is valuable as it keeps the space required for an index as small as possible and maximizes the speed of indexing or other data representation.

Meta data Categories and Use Cases

There are three main types of Meta data that are important in electronic discovery. The first two are attributes of file systems and applications and help identify who created, copied or modified documents. This capability helps to identify custody or ownership criteria for documents important to a case. The third type of Meta data is supplied by either human reviewers or analytic software processes.

File System or Repository Meta data

For example, the file system where documents are found has Meta data about who copied a file to the file system or when a file was created on a specific file repository. This category would include SharePoint Meta data, NTFS (Windows) file system Meta data and any kind of Meta data that is relevant to the repository storing a data item (when it was placed into the repository, how large it is, what Access Control Lists (ACLs) apply to control the viewing of the item, etc.). If a litigation support person was looking for files that were created on a file system during a specific time period, they would be interested in file-level Meta data. An ideal discovery solution always indexes the import path of any document it represents along with as many file system attribute fields as possible.

Application-level Meta data

The application (MS Word for example) that creates a document stores certain Meta data fields inside any documents it creates. This presents an additional type of Meta data that can be indexed and analyzed to identify documents with certain characteristics. Application Meta data contains fields like who the author of a document may be, when they created the file with the application (MS Word in this instance) or when the file was modified (inside the application). The ideal discovery solution would capture as many of these document-specific Meta data fields as possible to determine everything from authorship of the document to when it was last printed (depending on what application created the document).

User-supplied or "Analytic" Meta data

The last type of Meta data that the system can store for a user is "Analytic" Meta data. This is user or machine supplied Meta data. Even though final document tagging is done by an attorney within the final review stage of a legal discovery operation, other support personnel will mark or tag documents to indicate their status. Legal support personnel may need to mark or "tag" documents with labels identifying certain documents as "important" for some specific reason (the documents may qualify for "expert" review by a professional in a certain field of expertise for example). They may want to tag them so that a supervisor can review their work and decide that they meet certain criteria that qualify them to "move along" in the discovery process.

In addition to human review, a software analytic process can be run against a document collection and identify documents that are duplicate copies of one another in a large collection. An automatic process could generate tags (within the Analytic Meta data) indicating that certain documents are duplicates of a "master" document. If the master document was described as document "DOC000002345" then a tag such as "DUP_DOC1000002345" could describe all the documents that are duplicates of the master. These documents could then be identified quickly as redundant and they would not be passed along to attorneys for review. The system could retain the original copy of a duplicate document and mark or

remove the others so that attorneys would not have to read duplicates unnecessarily. The ideal discovery solution can run near-duplicate analysis and determine that certain documents meet a threshold of "similarity" to other documents, qualifying them as "versions" of an original document. Tags can then be automatically applied to the documents exhibiting these relationships so that they are identified for in-house counsel who may want to pass them along as data that outside counsel should review.

Analytic Meta data is the repository where an ideal platform can conveniently place both human and machine-assisted codes or tags that will streamline or aid review of documents in a later part of the process. Given that human review is very expensive machine-assisted "culling" of information can reduce costs dramatically. Many experts in the industry term this process as part of "assisted coding" or "predictive coding" of documents.

Analytic Processes

For purposes of this paper, "analytic processes" will refer to the following main functions within the ideal discovery solution:

1. Unsupervised Classification – some refer to this as "clustering" where documents are organized together into lists or folders with members exhibiting some level of semantic similarity to one another. The term unsupervised refers to the technique's ability to perform this semantic matching with no human supervision.
2. Supervised Classification – this refers to a capability where the product can take example content from a user and organize documents into lists using these examples as starting points or "seed" documents. The "best matches" are taken from among the candidate population of documents that are to be classified. The user can assign meaning to the seed clusters as they see fit; assign labels, etc. In the ideal solution a user can pick a number of documents as seeds, and specify an ordinal indicator of similarity that is a number between 0-1 that indicates a "threshold" of similarity that must be met for the candidate document to be placed on a seed list. Another form of the supervised classification is "search by document" where a user can select a single document as a "seed" and have it attract the most likely matches from the candidate list.
3. Near-duplicate analysis – this is very similar to supervised classification except that the system can take one "pivot" (example) document and compute all others within a relative "similarity distance" of it. Instead of organizing the document into a list of other semantically similar documents; candidates are marked as "near-duplicate" neighbors of a pivot should they fall within a range of similarity specified by a user. The documents are marked with "near-duplicate association" markers in the analytic Meta data repository as indicated above.
4. Email conversation analysis – this is where the ideal system identifies the email and instant messaging conversations that occur between parties. The parties and who sees a message is discernible through this type of analysis.
5. Different types of searching – simple keyword search, Boolean search, fuzzy search, proximate search are other types of search that are sometimes referred to as analytics within a product. An emerging technology that is more and more important to legal discovery is conceptual

searching, where concepts are computed among the members of documents and presented with the keyword results. Often conceptual searching is referred to in the context of conceptual mining which means a process that identifies concepts in documents that transcend keywords. Conceptual mining is often used to identify "latent" or immediately "unseen" words that are significant among a population of documents. These can often help a human reviewer identify what keywords should be included in a case and also to identify documents that the initial keyword searches did not include.

Virtual Index

For legal discovery purposes, a system needs to support building and searching the aforementioned three types of Meta data and must include support for analyzing and searching full document content as well. For analytics of certain kinds documents must be represented by special data structures that allow analysis (duplicate analysis, near-duplicate analysis, similarity comparisons to example content, etc.) to be undertaken. The system has to account for these at great scale.

This entire set of capabilities should appear (to a user of the system) to be possible across one "index". In the ideal system, these capabilities are encapsulated in one entity that will be referred to as: "the virtual index". It is referred to in this way because it supports various operations on multiple data representations and encapsulates these operations transparently to the user. The user should not know or care about the different repository or representations of the documents within the ideal system. The user should simply issue searches or ask for "similar documents" and get the results. The virtual index will abstract all of these details for a user.

Multi-site Support

The ideal system should support use cases "behind the corporate firewall" for analyzing and collecting data within local enterprise or client environments, and also support large data center deployments. The indices built within the enterprise environment should be "portable" so that they can be built in the enterprise environment and then be transported to the larger data center environment where all aspects of the case can be evaluated in one "virtual place". The idea of a virtual index supports this vision, as it allows local data sources to be analyzed at various remote locations and then any relevant files moved to a legal hold location at a central data center. The indices can be added to the central location along with any data that is copied for legal hold purposes.

In all cases it is ideal to have a platform that "connects" to data sources, reads in a copy of the documents stored within them, but leaves the original in place at its source location. Instead of moving the original document into the ideal system and duplicating the document and the storage required to maintain or analyze it, the documents can be represented by an index or some data structure that is generally more compact. The original documents do not have to be resident within the ideal system to be analyzed and referenced. Please see Figure Two (below) for an illustration of the ideal system in relation to data sources it represents.

It is important that documents do not have to be loaded and analyzed in "batches" and that the ideal system has the scale to represent vast numbers of documents within one single system. A system that

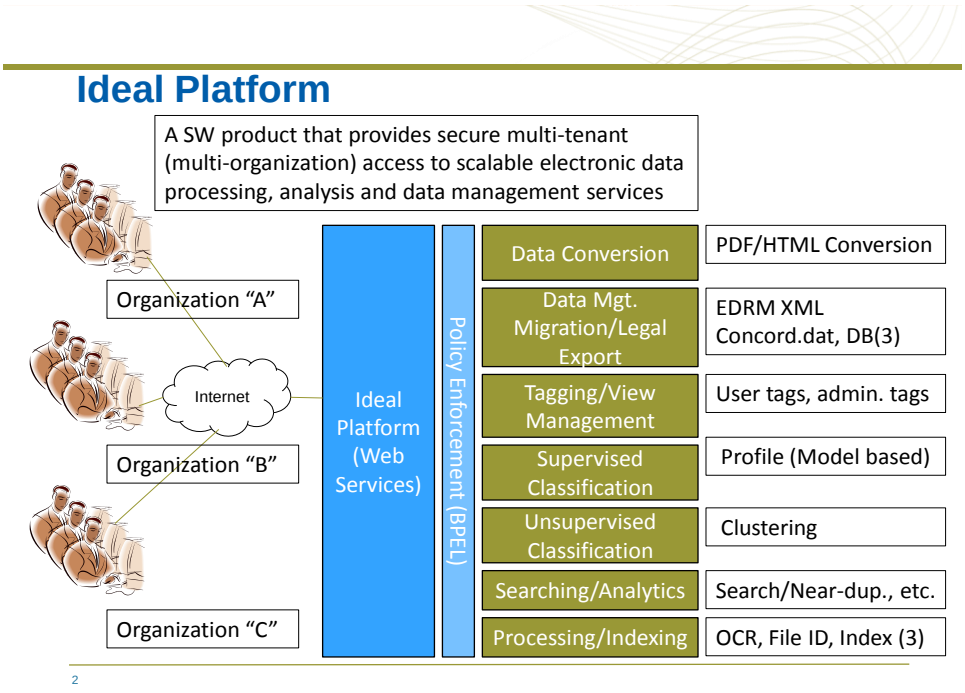
supports a set of analytic operations and scalable search is also an important feature of such a discovery platform. Having the ability to analyze new documents by comparing them analytically with examples already represented within the ideal discovery system is extremely important to solid ediscovery practices.

Key Architectural Attributes of an All-Inclusive Platform

An all-inclusive platform approach presents all of the capabilities shown above to the IT or legal review professional. The user can index data from locations within their data center or from sources as diverse as their SharePoint server "farm" their Exchange email server, NT file servers or large-scale NAS devices. The user can pick various levels of data representation based on the level of insight required for the task and the computational and storage burden acceptable to the reviewers. The user can then search for data that is relevant, select those results to "pass on" to other analytic processes (such as de-duplication and near-duplicate identification or email analysis) and then tag or otherwise mark the results.

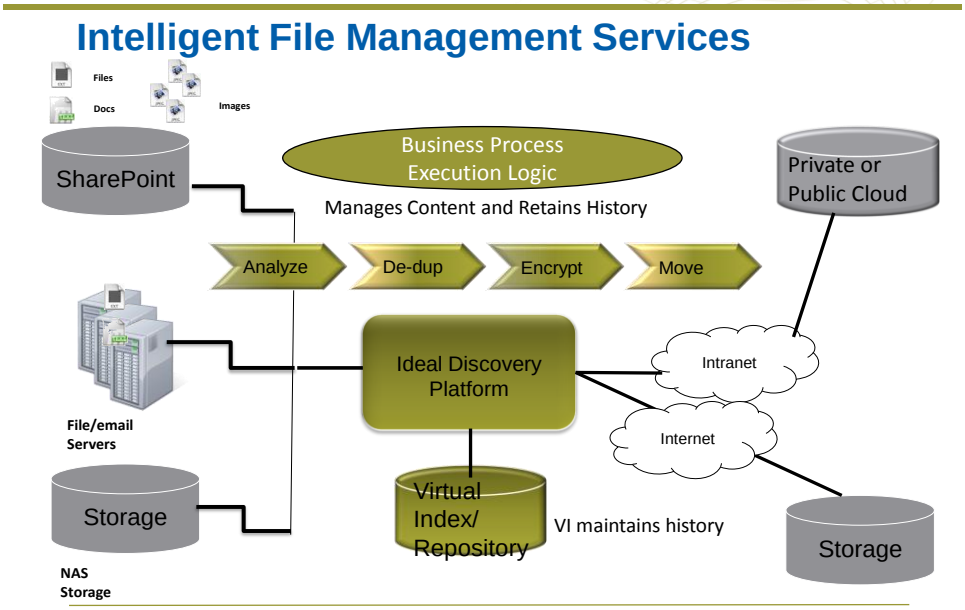
All of these capabilities should be available from a single console without the need for moving the data from one tool to another. Once the data is in the platform it can be identified, analyzed and marked according to the needs of the case. The important thing is that it can be managed with these processes at unprecedented scale. Please see Figure One (below) for an illustration of the ideal platform. The reader can quickly recognize that this is a product with a full suite of analytic and legal production capabilities. It is far beyond a single function product like a search engine. Please see Figure Two below for an illustration of how this platform could operate in the IT infrastructure among various repositories of data.

Figure One: Ideal Discovery Platform



The ideal discovery platform will perform all of the functions in the illustration above. The power of having all these capabilities in one platform is undeniable. Being able to process content (OCR, REGEX), index it, "cull" it down to a smaller size and then analyze it (remove duplicate material, perform NIST analysis, identify near-duplicate content, calculate email conversation "threads") all in one platform without having to move the content from one system to another eliminates labor and potential human error. Promoting efficiency in electronic discovery is a key component to success in legal review matters.

Figure Two: Intelligent File Analysis and Management



Scale of Indexing and Representation

An ideal discovery solution must have unprecedented scale. Scale is provided through superior use of physical computing resources but also through the segmenting of the various data resources into the virtual index components described previously.

Scale of Indexing, Search and Analytics [List of All Unique Terms in a Collection]

With the correct architecture hundreds of millions to billions of documents can be indexed and managed in a fraction of the time required for other solutions, and with a fraction of the hardware they require. One vendor, utilizing a unique grid-based (multi-server) architecture has demonstrated the indexing and preparation of a given 17.3 TB data set in less than a twenty-four hour period. This is possible due to two factors:

1. The platform's unique "Grid" architecture (see Figure Three)
2. The platform's unique "Virtual Indexing" architecture and technology (see Figure Five)

This platform can be deployed as a single server solution or in the large data center configurations shown in figure three below. The ability to expand as the customer needs to index and analyze more data in a given amount of time is made possible by the architecture. Certain software components of the architecture schedule activities on the analytic engine components shown in the diagram. These analytic engines "perform intensive work" (indexing, searching) and the controlling software requests them to perform the work to produce results for users. The controlling software is the "intelligence or brains" of the system and the analytic engines are the "brawn" of the system. As the user needs more processing power, more analytic engines can be employed within the "grid" to provide more processing and analytic power (the user is again referred to Figure Three)

Scale of Representation

This architecture also supports the representation of content in multiple ways so that the search, classification and other analytic operations available from the analytic engines can "work on" the data that has been processed. This means that the index is really a set of "managed components" which include:

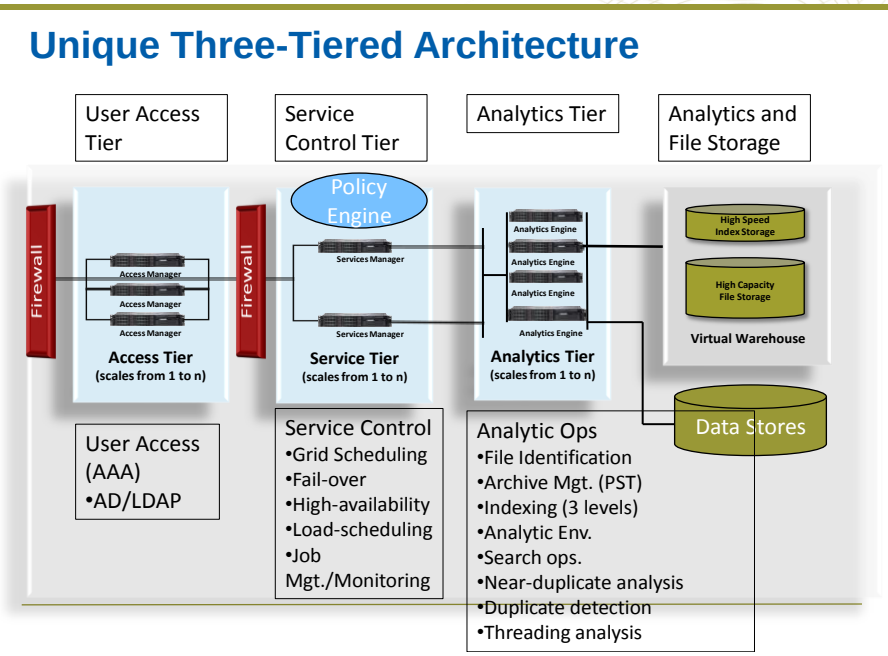
1. Meta data indices
2. Content indices
3. Analytic data structures
4. Analytic Meta data (tags, cluster groups, other associations)

All of these things are what is meant by "scale of representation"; the platform can represent content in multiple ways so that the appropriate level of search or analytics can be undertaken on the documents that are within a collection or corpus.

Speed and Scale of Indexing

A second aspect of scale is the speed with which data can be processed and made available for assessment. With a superior architecture an index can be presented for searching within hours. Other solutions require days if not months to build the content for a case into a searchable representation. The ability to build an index and get results in one or two days and have it done reliably allows case matters to be investigated rapidly and with fewer errors. The sooner a reviewer can determine what is relevant within the scope of discovery the sooner lawyers can begin making intelligent decisions about the case. This leads to better outcomes because the reviewers are not as rushed and because they have better analysis options than they would have with traditional methods. With the data prepared faster, organizations have time to perform search operations and then perform more complex analysis of data that will aid the reviewer later in the case.

Figure Three: Scalable Grid Architecture



As one can see from the illustration above, the data processing and search workload can be distributed over various machines in the "grid". The customer simply has to install and provision more "engines" to exist in the grid and the intelligent management layer of software will use these resources for processing, indexing and search operations. This allows the product to scale to handle unprecedented levels of documents and to process them in unprecedented timeframes. In Figure Eight (below) the search operation is illustrated as being distributed over the available grid processing power.

Virtual Indexing: a Key to Large Corpus Management

In addition to a distributed "grid-like" architecture, another key to managing large data sets is using the proper constructs to represent the data. As mentioned, this platform builds different representations of the data based on the needs of the analysis tasks that will be required for specific discovery activities. It ties them together in a logical set of components that is referred to as a "Virtual Index". This is necessary because the Meta data from files, the user-supplied Meta data from other reviewers, and analytically generated Meta data all must be searched as a single logical entity to make decisions about a given case.

A virtual index stores the various pieces of the logical index separately so that the Meta data can be built and searched separately for efficiency reasons, but also for scale purposes. A virtual index can be grown to an unprecedented size because it is "built up" from smaller more efficient components. Further, it can be transported from one location to another, and then "added in" to a matter as the user sees fit. Earlier in this document the example of remote office local collection with the data being transported with the appropriate indices to a data center. This is possible because of the virtual index. Such an index can also grow arbitrarily large. The virtual index component of software can "open" the parts of a virtual

index that matter to a case at a certain point in time. This makes searching more efficient and also it allows the virtual index to grow or shrink as necessary. Also, the “pieces” of a virtual index can be repaired if they become corrupt for some reason. The ideal system retains “manifests” of documents that comprise a given portion of the virtual index and from these the component indices can be rebuilt if necessary.

The user may want to just look at file system Meta data and characteristics of content stored within an enterprise. For that a straight forward file system Meta data index (basically POSIX-level Meta data) will satisfy the need. This type of index only requires about 4% of the original data size for storage. A full content index (on average) consumes between 25-30% of the original data size. The full-content index will require more storage than the Meta data variety of index, and it will take longer to build.

If the user needs to understand the application (MS Word, PDF) Meta data or that and the full content of documents for keyword search, they will be willing to wait for the extra processing (full content indexing) to complete and are likely willing to consume extra storage. If the user is not sure if all the available content meets the criteria that their search may require, they may want to use the POSIX Meta data indexing technique initially to identify what content should be fully indexed (before committing to extra time and storage resources).

One key aspect of the ideal system is that the Meta data index is separate and stands alone from the content index that supports it. The system presents one index for a given corpus of documents, but beneath this construct is at least a Meta data index. If a corpus is represented as a full content index, the corpus has a Meta data and a full content index component. The two indices are logically connected but physically separate. The virtual index software layer “binds” them together. Please see Figure Five for an illustration of a virtual index and its components. This virtual index approach makes the index more scalable; it allows it to be searched more rapidly and makes it resilient against potential corruption.

In addition to the full content inverted index construct, the corpus of documents can be further represented by analytic feature descriptors (where each “word” or “token” is represented as a feature of the document). These feature descriptors for single documents can be combined as “models” where complex relationships between the words or tokens are stored. These analytic descriptors are separate data structures that support document similarity operations, clustering and near-duplicate analysis. They do not depend upon the inverted index that is used for keyword searching; they are separate data structures and are used independently of the index.

Figure Four: Analytic Meta Data

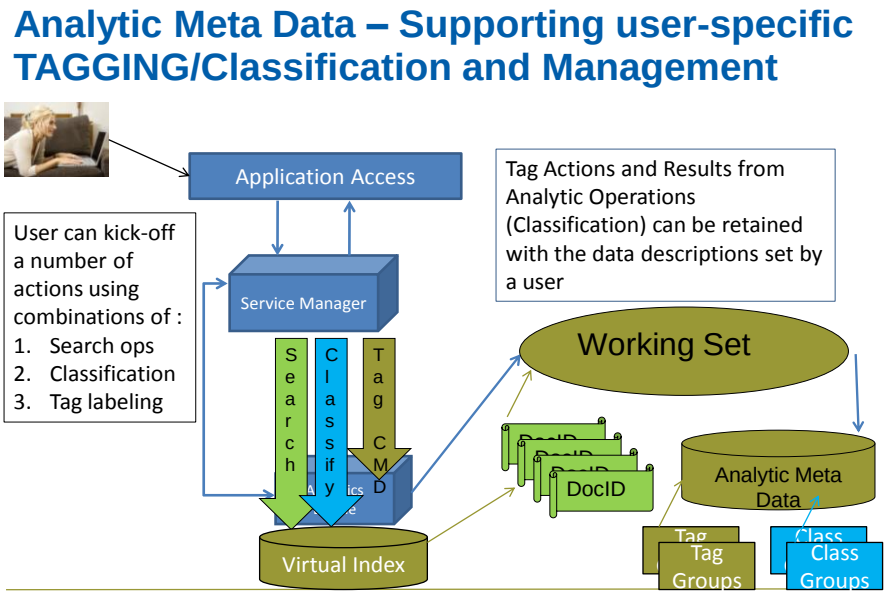


Figure Five: Virtual Index Illustration

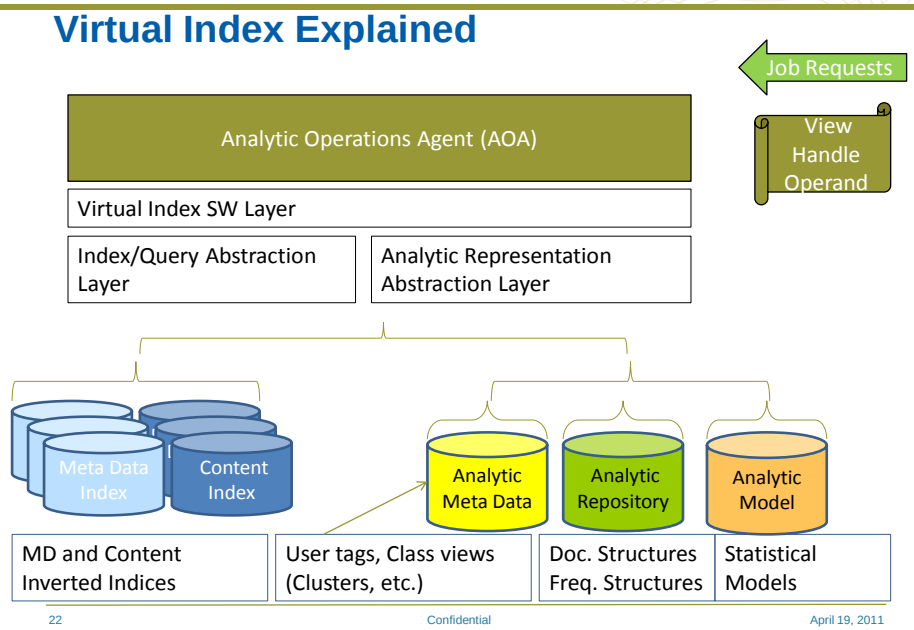


Figure Six: Virtual versus Monolithic indices

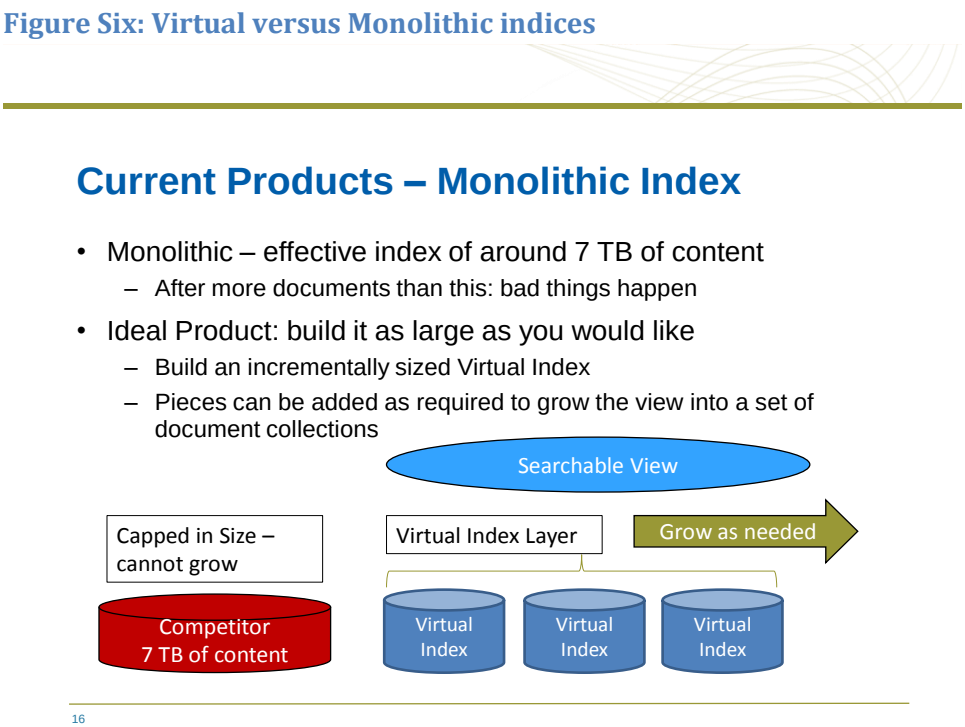


Figure Seven: Virtual Indexing in Action

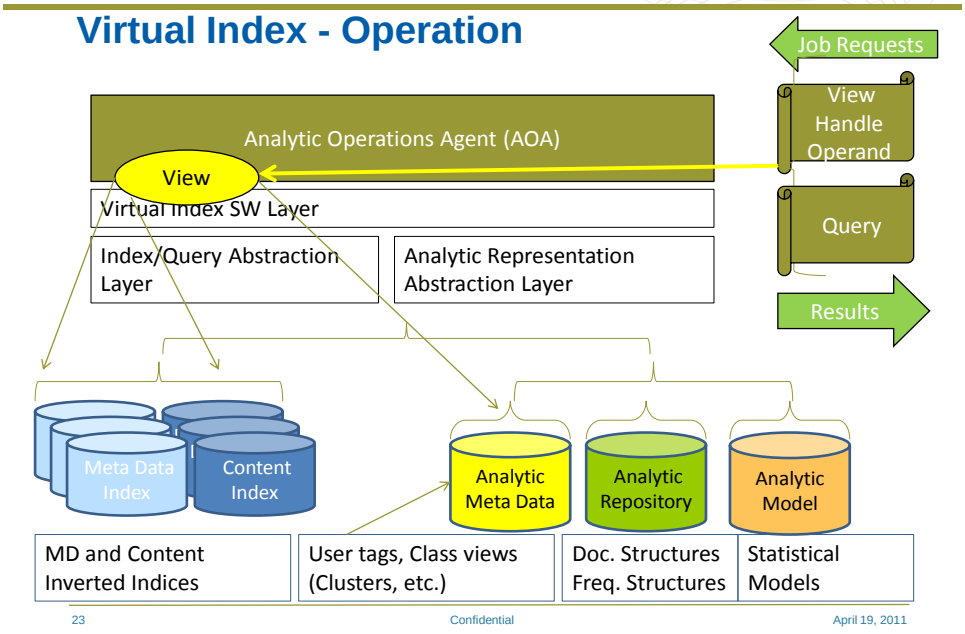
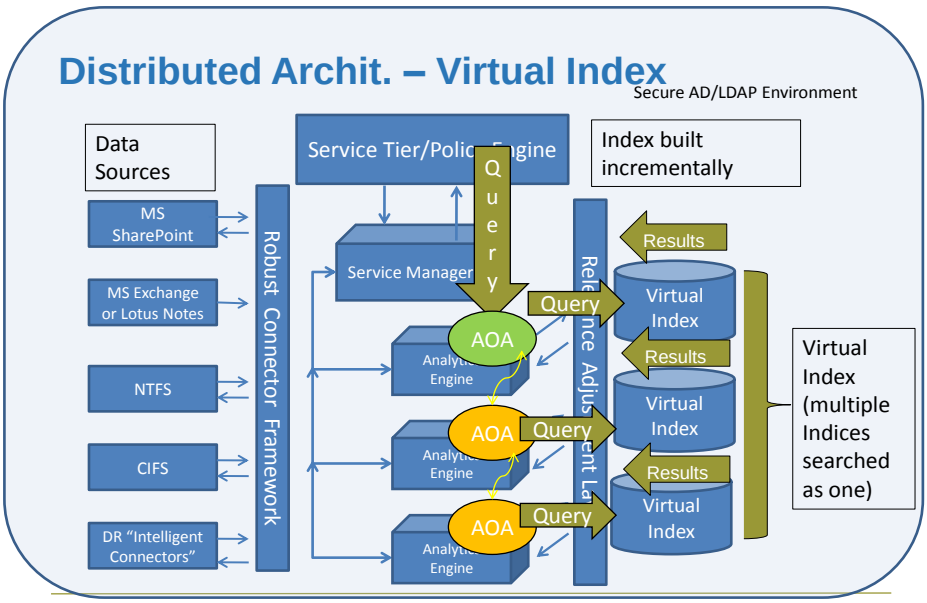


Figure Eight: Virtual Indexing plus Grid Architecture



Summary of Architectural Concepts

Now that we understand how the unique architecture of the ideal system solves several issues around Discovery, we can talk about some important types of processes which we will refer to as "analytics" and their importance to the overall process. The prior sections of this document explained how:

1. The grid architecture allows very large collections to be represented as indices in record time. Before this architecture became available, the Ediscovery process could not analyze the extremely large collections of documents that have become common in legal matters. These collections were either not analyzed, or they were analyzed in pieces, leading to human error and inconsistency of results.
2. The virtual index constructs let the user select various levels of index representation
    a. File Meta data only
    b. Application Meta data
    c. Full Content
    d. Analytic Descriptions (document feature attributes)
    e. Models and Profiles (example content in feature-attribute form)
3. The virtual index also lets the document collections that are represented grow to unprecedented size and still remain usable and efficient
    a. The virtual index lets the user add to the collection at any time as the virtual index is really a multi-index construct within the product
4. Monolithic Indices can be problematic
    a. Monolithic indices can grow in size and be very inefficient to search and manage

b. Monolithic indices can become corrupt at a certain size and become unusable
c. Monolithic indices can take long periods of time to construct in the first place

5. Virtual Indices Supply Several Key Advantages
    a. In a virtual index Meta data only searches work on smaller absolute indices and complete more rapidly
    b. In a virtual index Meta data and full content searches actually execute in parallel increasing efficiency and scale while providing results more rapidly than from classical monolithic indices
    c. Virtual indices can support similarity operations like “search by document” that expose relevant documents that are meaningful to a human reviewer
    d. Virtual Indices can be repaired efficiently without requiring entire document collections to be re-processed and re-represented.

Analytics

Analytic processes in EDiscovery present distinct advantages to human reviewers. In this section, analytic processes are described that can aid the legal review process. Discussion is presented about why they can be of help during that process. In addition, aspects of these approaches are presented and compared and advantages and disadvantages of each are explored. This is to give the reader a sense of how competing products in the discovery space compare. This is intended to help the reader value each and determine when one technique needs to be applied with others to be effective in a legal review process.

Major Categories of Analytic Processing

It is easy to become confused by all of the techniques that are available to aid human reviewers of electronic documents. These techniques fall into three main categories:

1. Unsupervised classification or categorization (often referred to as “clustering”)
2. Supervised classification or categorization
3. Specific types of analysis
    a. Near-duplicate analysis
    b. Duplicate identification or analysis
    c. Conversational analysis (who spoke to whom)

Unsupervised Classification

This is often referred to as “clustering” because one wants to form document groups that “belong together” because they “mean the same things”. The main idea behind this kind of analysis is that the human reviewer does not have to know anything about the data in advance. The reviewer can just “push the button” and find out what belongs where in a dataset and see folders or ordered lists of documents that are related somehow.

See Figure Twenty Two (below) for a screenshot of a product that identifies documents according to their similarity to one another. This particular system uses a series of algorithms to perform its work; but the end result is folders of documents that are related to one another. Close inspection of the diagram will show that the foreign language documents end up in the same containers or folders. The predominantly foreign language documents get grouped together in a folder that is labeled with foreign language “concepts” to make the review process more efficient. Other advantages of this technique will be explained in later sections of this document. This is an example of a multi-level algorithm that accounts for language differences. Some unsupervised classification algorithms do not account for the language differences of documents and they can produce results that appear “confusing” in some circumstances. This phenomenon will be discussed later in the document.

Most of these unsupervised techniques culminate in some kind of “conceptual” clustering or conceptual mining and analysis of the data they represent. As each specific technique is described in later sections

of this document the reader will be informed about how the technique relates to conceptual analysis of documents being analyzed.

Supervised Classification

Supervised classification means that a user "supervises" the process by providing at least some examples of documents that are similar to what they want the algorithm to find for them in a larger population of documents. These documents are usually put into what is called a "model" and they "attract" other documents that belong "closely" to them. Please see Figure Twenty Four for an illustration of supervised classification.

Examples of supervised approaches:

1. Seed-model "nearest neighbor example" type clustering.
2. Support Vector Machines (see reference [6]) – the user must supply "known positive" and "known negative" examples of documents that the system can use to "compute the differences" between for purposes of classifying new documents.
3. Bayesian Classifiers (see section below and reference [3]) – the user must supply "good" and "bad" examples so that the algorithm can compute a "prior" distribution that allows it to mark documents one way or the other.
4. Statistical Concept Identifiers that arrange documents based on the characteristics of words and topics in a set of "training data" (documents that have been selected from a larger population of documents but that have not been reviewed by a user)
5. Linguistic Part of Speech (POS) models where certain patterns of a specific language are noted within a linear classification model and new documents are "matched" against it based on their linguistic characteristics.

Specialized Analysis

There are specialized analytic techniques such as:

1. Near-duplicate detection (finding things that should be consider versions of other documents)
2. Email conversational analysis (threads of conversations between specific parties)

Mathematical Framework for Data Analysis

In the preceding discussion of the ideal architecture the concept of representing data as an index for searching or as a mathematical model for analysis was presented. This section contains a description of how data is represented mathematically. There are many ways to represent data for mathematical analysis; the technique being described below is one way. This is not intended to be an exhaustive review of all available text representation techniques; it is offered to help the reader visualize methods that are not the same as an inverted index that can be used as a basis for document analysis.

Basic Overview of Document Analysis Techniques

The basics of how documents are compared to one another relies on representing them as "units of information" with associated "information unit counts". This is intended to give the reader context to understand some of the terminology that follows. The goal of this is to support a mathematical process

that can analyze document contents: the “vector space model” [7]. The vector space model was developed by a team at Cornell University in the 1960’s [8] and implemented as a system for information retrieval (an early search engine). The “pros” and “cons” of the vector space model are discussed in the references, but since it is a good way to understand how to think about documents in an abstract and mathematical way it is explained initially. When we refer to this in general it will refer to the document-term representation model where documents can be thought of as vectors.

The term Vector Space Model would imply that in all cases we mean that the vectors are compared with cosine angular measurements after their “term frequency-inverse document frequency” attributes are computed. In the context below I discuss how that is possible but I don’t explain “tf-idf” in detail. The reader can consult [7] and [8] for information on computing similarity with tf-idf techniques. Furthermore, I am offering the model as an example of how documents can be represented for comparison. Other techniques than tf-idf used for cosine similarity comparisons use the vector concept so I want to make sure the reader understands the context in which this discussion is offered.

The Vector Space Model (VSM) is often referred to not just as a data representation technique but as a method of analysis. Some of the techniques mentioned in the sections that follow utilize this vector type model in some way (for representation; but their mathematical approaches are different). Not all of the techniques discussed use the vector space model, but it is presented to give the reader a grasp on how a document can be analyzed mathematically. Some form of vector is used in many cases to describe the document content. Some of the techniques just need some data structure that represents the words in a document and how often they occur. This is often constructed as a vector even if the vector space calculations are not used to analyze the data the vector represents.

Representing Text Documents for Mathematical Analysis – Vector Space Model

The "Vector Space Model" is a very well known structure within the field of information theory and analysis. It allows documents and their words or "tokens" to be represented along with their frequencies of occurrence. Documents are represented by a "document identifier" that the system uses to refer to it during analytic operations or so that it can be retrieved for a user. The overall combination of document identifier and the token frequency information is referred to as a "document descriptor" because it represents the information with the document and provides a "handle" to use to grab the document when necessary.

Figure Nine: Vector Document term Frequency Structures

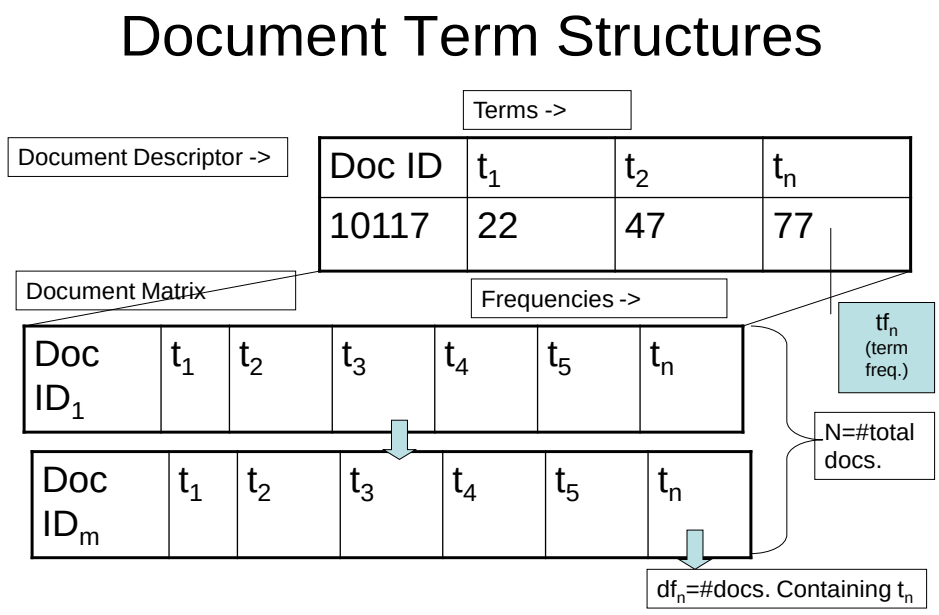
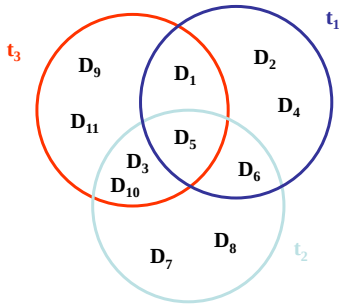


Figure Ten: Documents in a Matrix

Vector Space Documents Matrix Representation and Queries

docs	*t1*	*t2*	*t3*	RSV=Q.Di
D1	1	0	1	4
D2	1	0	0	1
D3	0	1	1	5
D4	1	0	0	1
D5	1	1	1	6
D6	1	1	0	3
D7	0	1	0	2
D8	0	1	0	2
D9	0	0	1	3
D10	0	1	1	5
D11	1	0	1	3
Q	1	2	3	
	q1	*q2*	*q3*	

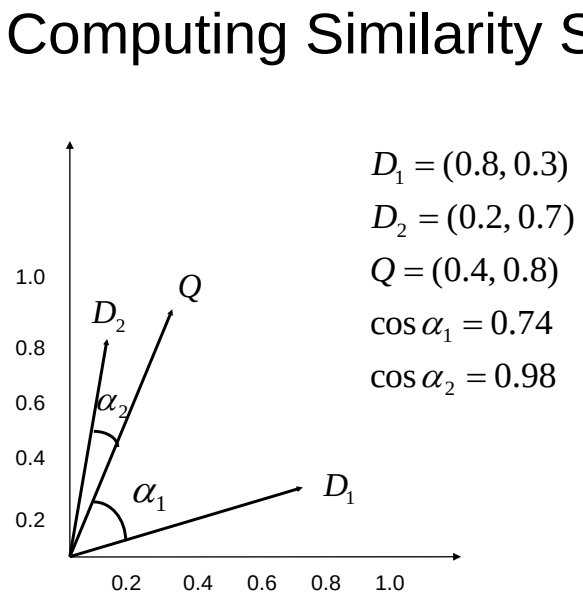


Notice that when documents are represented as document-term structures the documents are like "rows" in a matrix. The "columns" of the matrix are the terms, and the columns can be the frequencies of the given terms that are found to occur in the documents. The term positions can be fixed (per term) and labeled with some integer with the actual string of the word/token being kept in a separate dictionary or some other means can be used to "keep track" of what the terms mean. The representation of a document will be the document being an entire row of terms with the columns representing the frequency of occurrence of any given term within a specific row or document.

Comparing Documents to One Another or Queries

With the vector-space model of vector comparison, each document is treated as a "vector" in a dimensional space. The number of dimensions equals the number of terms in the largest document. If a document has a given term in a row of the matrix, the value in the matrix is equal to that document's frequency for the given term. If the document does not have that given term, the column in the row of a given document is zero. To compare two documents, a "similarity calculation" is undertaken and a "score" is computed between the documents. The score represents the cosine of the angle between the two documents, or their "distance apart" in the "n-dimensional vector space". This can be visualized in two-dimensions below. A query can be represented as a document so a query entered by a user can be compared to documents and the closest ones can be retrieved as results.

Figure Eleven: Cosine Similarity



The basics of the vector space model are that the cosine angle can be computed between any two vectors in the document-term matrix [7]. This number is guaranteed to be between zero and one and it shows that a document is identical to another document (score equals one) or the document has a zero score (nothing in common with the reference document) or something in between. The closer that the score is to the number one, the more similar two documents are in the "vector space" or "semantic space" of the documents. This model gives the reviewer some idea of how similar two documents are. It is useful in this respect; but it has some limitations. The reader can review the references for more detail

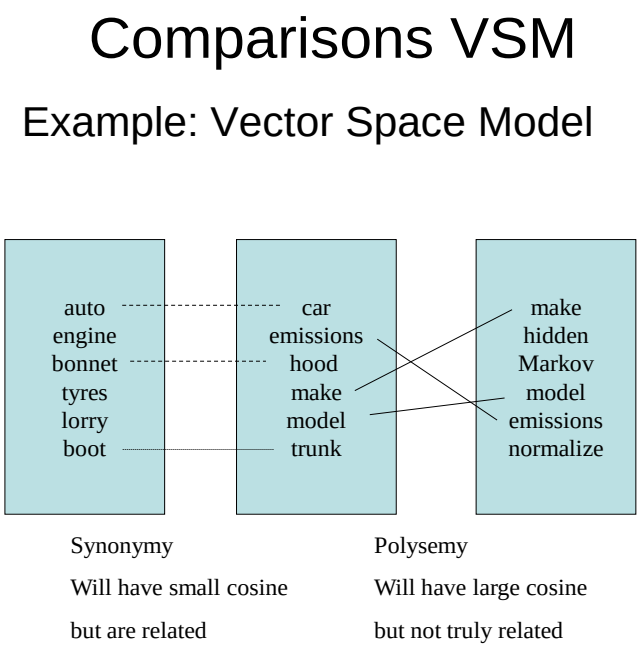
on the mathematics, but the basic idea is that documents are: 1) the same; 2) totally unrelated; 3) somewhere in between.

Problems with Vector Space Model (VSM)

The problems that arose using the vector space model included:

- synonymy: many ways to refer to the same object, e.g. car and automobile
- polysemy: most words have more than one distinct meaning, e.g. model, python, chip
- Vectors are still sparse and there is a lot of extra computation involved with analyzing them

Figure Twelve – Illustrative Behavior Vector Space Model



As can be seen above, the VSM puts things together that have the same (literal) words. It is efficient to compute and gives an intuitive basis for understanding what is literally similar. What it does not help with is in finding documents that "mean" the same things. In the example above, one document with "car" and another with "auto" would not be grouped together using this technique. This is because the technique cannot account for synonyms or polyesters (these are explained in [5] within the references); polyesters are words that have more than one meaning ("Java meaning coffee" and "Java meaning the island of Java", or "Java the programming language").

There are ways to improve the behavior of the vector-space model and it is still used and proves very useful for many operations in data analysis. For conceptual analysis of data however, there are other methods that can be used that don't suffer from these drawbacks.

VSM: Historical Significance

The VSM as one of the first techniques to model documents in a mathematical context and present them in a fashion that is conducive to analytic review. It is not perfect, but it represents a lot of value to folks trying to find similar documents and it paved the way for researchers to use a common model for thinking about document analysis. The techniques that were subsequently put forward used this representation method but applied very different mathematical techniques to the matrix model of viewing document collections.

There are ways to improve the behavior of the vector-space model and it is still used and proves very useful for many operations in data analysis. For conceptual analysis of data however, there are other methods that can be used that don't suffer from the drawbacks of VSM. These techniques perform dimensionality reduction on the data sets at hand, and expose "latent relationships" in data that are hard to find otherwise. Before moving on to techniques, some discussion of reducing the dimensionality of data sets is presented.

Some More Basics

Before we discuss the various techniques at our disposal for legal discovery analytics; we have to discuss a general concept in language theory: "dimensionality". Dimensionality is the number of words or the number of things that we have to account for in a language model.

Analyzing Text Documents – "The Curse of Dimensionality"

The problem with text is that with most languages there are so many words to choose from. Documents vary in their vocabulary so much that it is hard to build one mathematical model that contains all the possibilities of what a document might "mean". Language researchers and computer scientists refer to this problem as "the curse of dimensionality" and many information analysis approaches seek to reduce the number of dimensions (words) that their models have to contain.

In the matrix above, if this represented a "real" collection of documents, the columns of the matrix would be much more numerous and for many of the documents the columns would not have an entry. This is what is meant by "sparse data" within a "document-term matrix". Many approaches are aimed at identifying what words in a document or set of documents represent "enough" of the total so that others can be ignored. Information theorists refer to this as removing "noisy data" from the document model. This concept revolves around choosing enough of the attributes (words) within the documents to yield a meaningful representation of their content. Other techniques are used to actually remove words from the documents before they are analyzed.

Stop Word Removal

Certain words (such as "a", "and", "of") that are not deemed "descriptive" in the English language can be removed from a document to eliminate the number of dimensions that an algorithm needs to consider. This may be helpful in some contexts and with some algorithms; it does reduce the numbers of dimensions that need to be considered by an analysis algorithm. When these are removed it can be hard to determine specific phrases that might carry meaning to a legal reviewer however. A search engine that can find phrases may not consider the difference between two documents with similar phrases:

Document #1:“We agree on the specific language outlined below…”

And:

Document #2: “We agree to pursue a process where we agree on a specific language to describe….”

In these two documents many search engines would produce both documents; each with clearly different meanings in response to a phrase search of: “agree on the specific language”. This may not be a problem to a reviewer because both documents are likely to be returned; but the reviewer will have to read both documents and discard the one that is not specific enough for the case at hand. In this instance, stop word removal would yield results that are less specific than a reviewer might want. The searches with this type of index may produce more documents, but the cost will be that they may not be as specific to the topic at hand.

Stemming of Language

With most languages, there are ways to find the “stems” or “root meanings” of many words through an algorithm pioneered by Martin Porter [9] that has been named: “Porter Stemming”. This technique has been used widely and is often referred to simply as: “stemming”. Any serious language theorist recognizes the term “Porter Stemming”. The algorithms were initially released for English language analysis but have been extended for many other languages. See the reference (again [9]) for more discussion of the techniques and the languages supported.

The idea with porter stemming is to reduce words with suffix morphologies to their “root” meaning. The root of “choosing” is “choose” and would show up in some stemmers as: “choos”. The roots of many common words can change after stemming to common “roots”:

```
alter
alteration
altered

become:

alter
alter
alter
```

This reduces the number of tokens that a document has to account for and the argument for using this technique is that the meaning of the words is “about the same” so the corresponding behavior this induces in the mathematics will not be deleterious to any given analysis technique.

The theory behind using stemming for search is that more documents of “about the same meaning” will be produced for a given query. In a legal review context documents that are not specific to a query could be returned with stemmed collections. This is similar to the situation that could exist when stop-words are removed from a collection. For analytic approaches, the same thing can occur. The algorithms that group documents together could produce results that are less specific than might be desired by a human reviewer.

For analytic approaches, the designer of an algorithm must consider this trade-off between precision and recall. The benefits of having fewer things to keep track of in the model may outweigh any lack of clarity around usage that the token suffixes may have conveyed. In a legal discovery "clustering" context this may not be true (as we will discuss), but stemming is an important attribute of a collection of documents that should be considered when preparing documents for legal review purposes. It can help immensely and it can make other things less specific (which it was designed to do) than one might want for a legal discovery application. The designer of the analysis system should consider how the documents need to be prepared for the optimal performance inside the algorithms the system will implement.

Higher-Order Mathematical Techniques

To this point, we have seen how a legal discovery platform must include many different pieces of functionality and respect both Meta data and full content search at great scale. We have also seen how analytics can be important to legal discovery professionals. We have set the framework for how to represent documents in a way that allows mathematical operations to be defined on abstract representations of their content.

We reviewed the vector space model and how document matrices have many "dimensions" that impact analytical performance for legal review purposes. We have discussed how to reduce dimensions by removing certain words from a collection or by reducing certain words to their "root forms" so that they can be considered more generally with fewer burdens being placed on the modeling technique. These techniques can reduce the specificity of results returned by the system. This may or may not be acceptable for a legal review application. For conceptual analysis of data, there are other methods that can be used that perform dimensionality reduction on the data sets at hand in a different manner.

One of the first solutions to this problem that was proposed was Latent Semantic Indexing (or Analysis) by a team or researchers at Bell Laboratories in 1988. Before these are explored, some of the commonly used techniques are listed and discussed. This is not intended to be an exhaustive review of every technique available for language analysis. It is not a critique of any technique or vendor implementation. This is a discussion of some common techniques that have been used within legal discovery products and presents a "pro" and "con" set of considerations for the reader.

Commonly Used NLP Techniques

Within legal discovery, there are some NLP techniques that have become commonly known within the industry. These have been championed by vendors who have had success in providing them as pieces of various edicovery products. These are:

1. Latent Semantic Analysis/ Indexing (LSA/LSI) – this is an unsupervised classification technique used in a popular review platform and some other products
2. Probabilistic Latent Semantic Indexing or Analysis (PLSI; sometimes referred to as PLSA) – this is a supervised learning technique that has been implemented within search engine products
3. Bayesian Modeling (this is described below; the term "Bayesian" is commonly understood for SPAM filtering and other knowledge based products)

4. Discrete Finite Language Models (companies with these technologies have used linguists to build a "rules based" engine of some sort based on the "Parts of Speech" found in a text collection) these are included as "linguistic models and algorithms" that they use to help find keywords for search and to "understand" collections. These probably are useful in some contexts; generally these are specific to a given language and will not provide much value to other languages without tuning by the authors of the model.

Techniques Discussed/Analyzed

Each of these techniques will be discussed briefly in the context of their use within legal review. All of these are of course useful in the appropriate context. Their usefulness in certain situations and what needs to be added to them to make them an integral part of the legal review process is noted below. Their behavior at a certain scale can become problematic for each technique; this will be discussed below.

Latent Semantic Analysis

Latent Semantic Analysis (sometimes referred to as Latent Semantic Indexing or "LSI") was invented by a team of researchers at Bell Laboratories in the late 1980's. It uses principles of linear algebra to find the "Singular Value Decomposition" (see reference [10]) of a matrix which represents the sets of independent vectors within the matrix that exhibit the best correlations between term members of the documents it represents. Notice that documents represented as "vectors" in a matrix make this technique available in the same way that vector space calculations are (as described earlier) available for VSM similarity operations. With LSI/LSA the terms that emerge from the document-term matrix are considered "topics" that relate to the documents within the matrix. These topics are referred to as the "k" most prevalent "topics" or words in the matrix of document terms.

LSA – "The Math"

From reference [11] (Wikipedia page on LSI):

"A rank-reduced, Singular Value Decomposition is performed on the matrix to determine patterns in the relationships between the terms and concepts contained in the text. The SVD forms the foundation for LSI.[15] It computes the term and document vector spaces by transforming the single term-frequency matrix, *A*, into three other matrices— a term-concept vector matrix, *T*, a singular values matrix, *S*, and a concept-document vector matrix, *D*, which satisfy the following relations:

$A = TSD^T$

$$T^T T = D^T \quad D = I_r \quad TT^T = I_m \quad DD^T = I_n$$

$$S_{1,1} \geq S_{2,2} \geq \ldots \geq S_{r,r} > 0 \quad S_{i,j} = 0 \text{ where } i \neq j$$

In the formula, **A**, is the supplied ***m*** by ***n*** weighted matrix of term frequencies in a collection of text where ***m*** is the number of unique terms, and ***n*** is the number of documents. **T** is a computed ***m*** by ***r*** matrix of term vectors where ***r*** is the rank of **A**—a measure of its unique dimensions ≤

min(*m*,*n*). **S** is a computed ***r*** by ***r*** diagonal matrix of decreasing singular values, and **D** is a computed ***n*** by ***r*** matrix of document vectors.

The LSI modification to a standard SVD is to reduce the rank or truncate the singular value matrix **S** to size ***k*** « ***r***, typically on the order of a ***k*** in the range of 100 to 300 dimensions, effectively reducing the term and document vector matrix sizes to ***m*** by ***k*** and ***n*** by ***k*** respectively. The SVD operation, along with this reduction, has the effect of preserving the most important semantic information in the text while reducing noise and other undesirable artifacts of the original space of **A**. This reduced set of matrices is often denoted with a modified formula such as:

$$\mathbf{A} \approx \mathbf{A}_k = \mathbf{T}_k \mathbf{S}_k \mathbf{D}_k^T$$

Efficient LSI algorithms only compute the first ***k*** singular values and term and document vectors as opposed to computing a full SVD and then truncating it."

This technique lets the algorithm designer select a default number of topics which will be "of interest" to them (the default value is usually between 150-300 topics). There is research to indicate that around 100-150 topics is the "best" or "optimum" value to configure when using LSA. This sparks debate among language theorists but has been discussed in other documents (see reference [4] and [11]).

The topics generated via LSA SVD decomposition are referred to as the "k-dimensional topic space" within the new matrix. This is because there are k (100-150-300) topics or terms that now "matter" (instead of the thousands of individual terms in a set of documents before the dimensionality reduction has occurred). So the original matrix that could have contained thousands of unique terms is now represented by a much smaller matrix with terms that are highly correlated with one another. Figure Thirteen outlines some of the mathematical concepts that apply with Latent Semantic Analysis.

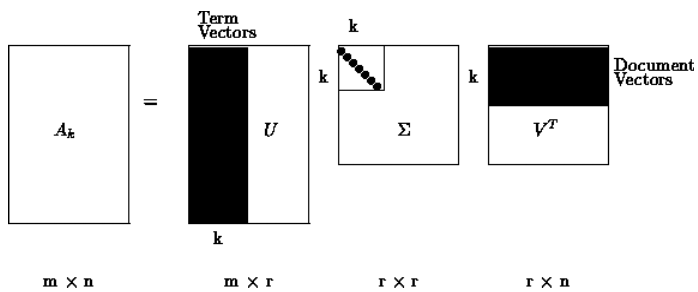
Figure Thirteen: LSA Matrix Illustrations

LSA: One of the First Solution

- Singular Value Decomposition

  {A}={U}{S}{V}^{T}

- V and U forms an orthonormal basis for input and output space: A*A, AA*



In the diagram it can be seen that the large matrix has been “reduced” to the smaller dimensional area and “important” terms are represented in the matrix. What the “mathematics removed” were topics or terms that did not appear strongly in relation to the terms that “survived” the operations that reduced the larger matrix. So it seems like this is a great idea (it was; it just is not perfect).

LSA Practical Benefits

To help the reader see the benefits of LSI, and how it can find correlations in data, an actual example is provided from a blog maintained by Alex Thomo [mailto:thomo@cs.uvic.ca]. This example was used in an earlier section of the paper to illustrate how LSA as a technique is very valuable. Here we delve into it a bit more and explain its “pros” and “cons”:

>>>

> An Example
>
> Suppose we have the following set of five documents
>
> d1 : Romeo and Juliet.
> d2 : Juliet: O happy dagger!
> d3 : Romeo died by dagger.

> d4 : "Live free or die", that's the New-Hampshire's motto.
> d5 : Did you know, New-Hampshire is in New-England.
>
> and search query: dies, dagger.
>
> A classical IR system would rank d3 to be the top of the list since it
> contains both dies, dagger. Then, d2 and d4 would follow, each containing
> a word of the query.
>
> However, what about d1 and d5? Should they be returned as possibly
> interesting results to this query? A classical IR system will not return
> them at all. However (as humans) we know that d1 is quite related to the
> query. On the other hand, d5 is not so much related to the query. Thus, we
> would like d1 but not d5, or differently said, we want d1 to be ranked
> higher than d5.
>
> The question is: Can the machine deduce this? The answer is yes, LSA does
> exactly that. In this example, LSA will be able to see that term dagger is
> related to d1 because it occurs together with the d1's terms Romeo and
> Juliet, in d2 and d3, respectively.
>
> Also, term dies is related to d1 and d5 because it occurs together with
> the d1's term Romeo and d5's term New-Hampshire in d3 and d4,
> respectively.
>
> LSA will also weigh properly the discovered connections; d1 more is
> related to the query than d5 since d1 is "doubly" connected to dagger
> through Romeo and Juliet, and also connected to die through Romeo, whereas
> d5 has only a single connection to the query through New-Hampshire.

<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<

>End Example from Blog

The example shows how the LSA technique can "link" together the concepts across the document collection. Even though the query has nothing to do with New Hampshire; the state motto: "live free or die" associates the query with the state. The power of the latent technique is obvious; it also can lead to obfuscation as we will also see in the next section.

The Power

LSA provides a set of concepts that were "latent" or unobserved in the data from the original document matrix. Due to the mathematical technique of computing linearly independent vectors that have pronounced term correlations; things that "belong together" show up because they relate to common linking words in certain ways. In a document collection, terms such as "astronaut" will be paired with "rocket" and "space" and "travel" or "expeditions". Terms such as "cosmonaut" will be related to the same terms. The astronaut and cosmonaut term ascendancies are good examples of the benefits with

LSA. A human reviewer may not have thought about cosmonaut as a possibility for a keyword search along with astronaut but after LSA reveals it as a latent concept the reviewer can include it in the keyword list of a given matter.

For legal reviewers it is valuable to find latent terms that are non-obvious that can be included in with obvious keyword selections. Other relationships in the data can be seen so that the legal reviewer can consider unseen aspects of document collections for building their legal strategies. If the document set is appropriately sized, the lawyer can receive "good ideas" from LSA operations that are run on data.

Another benefit to the technique is that new documents that arrive after LSA has been performed can by "folded in" to an existing reduced matrix (with a vector multiplication operation). This is often necessary as documents for a given case are often found as part of an iterative process where prior review leads to a widening scope of collection. The technique can also have applicability across languages as it may identify correlations in documents that are similar regardless of language [11]. This is not universally true for all languages, but it can be seen in some instances. There can be drawbacks to the LSA approach however.

The Problems: LSA Limitations

For large document collections, the computational time to reduce the matrix to its k important dimensions is very great. On large document collections (100,000 – 200,000 documents) computing SVD's can take a month or more; even with large capable servers equipped with large memory configurations. The technique does not distribute well over a large number of computers to allow the computational burden to be shared.

Even if the computational burden is acceptable, the technique is difficult to use after a certain number of documents because the data seems to put too many things in too few buckets. The terms that it seems to correlate don't always seem to make sense to human reviewers. This is due to a problem statisticians call "over-fitting". Too many wide-ranging topics begin to show up in the reduced matrix. They are related somehow, but it is not clear why.

Some good examples: good correlations occur where the terms "astronaut" and "cosmonaut" are paired with "rocket" and "space" and "travel". This all makes sense, cosmonauts and astronauts engage in space travel. But also included in the matrix are documents containing travel documentary reviews of the Sahara desert, "camels", "Bedouins" and "Lawrence" and "Arabia". These don't seem at all related to the documents about space travel. This occurs because the correlation of these topics with the ones about space travel relates to long journeys over harsh dry regions with little water, harsh temperatures and environments forbidding or deadly to humans. This over-fitting occurs as more documents with more topics exhibit these correlative effects. Soon there are too many associations to be crisp and logical to a human reviewer. Even in the example with just a few documents, it does not make sense that "New Hampshire" was introduced into the search results as "relevant" when the terms of the query were: "dies" and "dagger". If this were a murder case it would not have made sense to drag in documents that are about the state of New Hampshire.

So the dimensionality of the matrix was reduced with LSA and there are fewer things for a human to consider, but its discerning power was reduced as well. The results that emerge from the technique are confusing and do not lead to crisp conclusions around what the document population "represents". The technique is a purely mathematical one; there is no syntactic knowledge imparted to the model to check for consistencies with language usage. So it is clear that LSA is important, helpful under certain circumstances and that also it can be a bit confusing. Across a large population of documents it can take a long time to compute the relationships between documents and the terms they contain and the results of all that computation can end up being confusing.

Probabilistic Latent Semantic Analysis

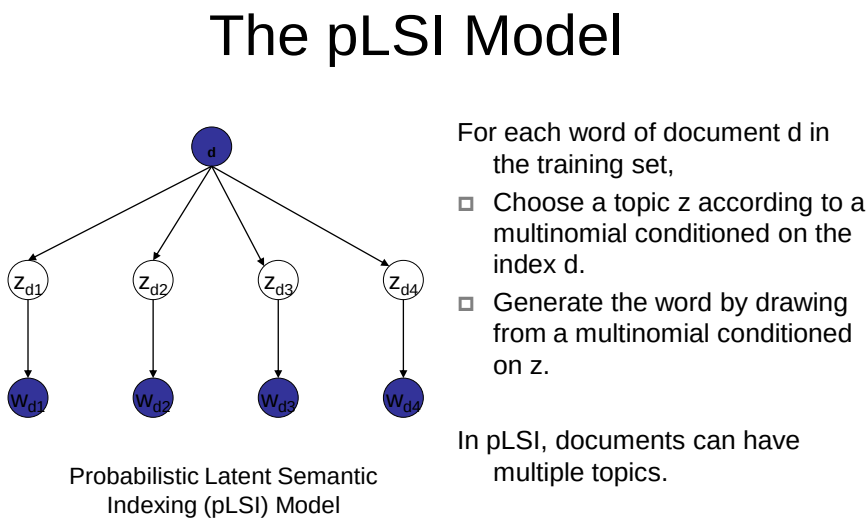
Because of the discernment issue with LSA and as a result of other researchers looking at the problem of conceptual mining in a new way, Probabilistic Latent Semantic Analysis, or PLSA was invented. The reader is referred to [2] in the references section for a full discussion of the technique, but it is built on a foundation from statistics where co-occurrences of words and documents are modeled as a mixture of conditionally independent multinomial distributions.

Instead of using linear algebra to reduce a matrix, the PLSA technique looks at how often a certain topic occurs along with a certain word. The following formula is from [2] and it basically says that for a given class of documents, the word "w" occurs at a certain frequency or with a certain probability along with the topic "z". This is found by iterating over a training set of documents and finding the highest correlations in the documents that contain both w and z. This is what they mean by a "multinomial distribution" within the documents of the collection. This technique was invented by Thomas Hoffman at Brown University (and others) and is referenced in [13].

$$P(w,d) = \sum P(c)P(d \mid c)P(w \mid c) = P(d) \sum P(c \mid d)P(w \mid c)$$

This may be easier to visualize with an illustration. It can be seen that the user picks a representative set of documents and then the PLSA software finds the highest valued topics for each word. This is accomplished with what is called an "Expectation-Maximization" algorithm that maximizes the "logarithmic likelihood" that topic z occurs with word w for a given word document combination.

Figure Fourteen: PLSA Illustrated



Benefits to PLSA

The benefits that PLSA has are that correlations can be found with a statistical basis from a document population. One can say that there is a definite and certain "likelihood" that certain documents contain certain optics. Each document can contain multiple topics as well.

Drawbacks to PLSA

This technique represents the topics among a set of training documents, but unlike LSA it does not have a natural way of fitting new documents into an existing set of documents. It also is computationally intensive and must be run on a population of documents selected by the user. Unlike LSA, it is a supervised classification method; it relies on a set of documents identified by a user. If the population of documents selected by the user is not representative of the entire collection of documents, then comparisons to documents that have been analyzed previously are not necessarily valid. One cannot take into account any prior knowledge of the statistics underlying a new unclassified data set.

PLSA (like LSA) also suffers from over-fitting. With PLSA several hand-selected parameters have to be configured to allow it to perform acceptably. If these "tempering factors" are not set correctly, the same issues with ambiguous topic identification can be seen with PLSA (as with LSA). Given that most users of data analysis products don't understand the impacts of hand-tuning parameters, let alone the techniques being used (the mathematics involved) this concept of setting parameters in a product is impractical at best. Therefore, PLSA is a statistically based and mathematically defensible solution for concept discovery and search within a legal discovery product, but it can be quite complex to tune and

maintain. It is likely a very difficult technique to explain to a judge when a lawyer has to explain how PLSA might have been used to select terms for search purposes.

Problems with Both LSA and PLSA

With both PLSA and LSA, a product incorporating these must still provide an inverted index and basic keyword search capability. Therefore, if one just implemented PLSA or LSA, the problem of providing a scalable keyword indexing and search capability would still exist for legal discovery users. All of the problems that were presented in the first part of this document still exist with platforms supporting these two analytic techniques.

Bayesian Classifiers

Bayesian Classifiers are well known for their work in the area of SPAM detection and elimination. Basically they find "good" examples and "bad" examples of data and compare new messages to these to determine if a new message should be classified as one or the other. The mathematics behind this kind of technology is discussed in [3] and is based on "Bayes Theorem" of conditional probability:

"Bayes Theorem basically says that a document probability of belonging to a certain class ("C" in the equation below) is conditional on certain features within the document. Bayesian theory relies on the fact that these are all independent of one another. This "prior" probability is learned from training data.

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\ p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}.$$

From [3]: "in plain English the above equation can be written as":

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

These can be used in a legal discovery context and some companies employ these types of technologies in their products. These kinds of classifiers are useful, they just have to be trained to accomplish their work, and this requires a human to perform this prior classification.

Bayesian Benefits

When properly trained, they work quite well. They are surprisingly efficient in certain cases. Like all tools, they are good at certain "jobs".

Bayesian Classifier Drawbacks/Limitations

They sometimes have no idea what to do with unseen data; if there is no example to guide them, they can make "bad choices". They can take skilled humans to collect the data for the "models" that they need to be effective. A lot of times this is not possible and can lead to unproductive behavior.

Natural Language Models: AKA Language Modeling

There are products that claim to have "proprietary algorithms" where "linguists" construct classifiers based on part of speech tagging (POS tagging), or specific dictionary based approaches that they feel "model" language. These often require professional services from the same companies that sell the software implementing the models. This is because the linguist who constructed the model often has to explain it to users. In a legal setting these approaches often require the linguist to become an expert witness if the model results come under scrutiny. These are not stand-alone software tools that one can run at the outset of a legal matter to "get some ideas" about the electronic information available for a case.

These models often require hand-tuning of the models given an initial keyword set produced by attorneys who have some initial knowledge about a case and are therefore not tools to expose meaning in language innately. They are more like "downstream" language classifiers that help identify documents in a large collection that meet some well understood semantic criteria established by the keyword analysis.

There are other products that use a combination of dictionaries and language heuristics to suggest synonyms and polyesters [5] that an attorney could use for keyword searches given a well-understood topic or initial list of keywords. These also may require that an expert explain some of the results if there is a dispute over the keywords it may suggest.

Drawbacks to Linguistic Language Models

The drawbacks to these methods include:

1. They often require hand-tuning and are not general software packages that can organize and classify data
2. They often require professional services and expert witness defense
3. They are very language specific (English, French, etc.) and do not scale across multi-lingual data sets

Other Specialized Techniques

Near-duplicate Analysis

Often versions of documents that are measured to be within "some similarity measure" of reference or example documents are very useful to identify. Knowing that a certain document has been edited in a certain place and in a certain way can be very useful to a legal reviewer. Knowing when this is done on a timeline basis is again a very crucial piece of many legal cases. Near-duplicate data identification products perform this kind of analysis.

For two documents:

- A near-duplicate identification product would build efficient data structures to compare two documents:
  - "Mary had a little lamb" – document #1

- "Mary had a little white lamb" – document #2

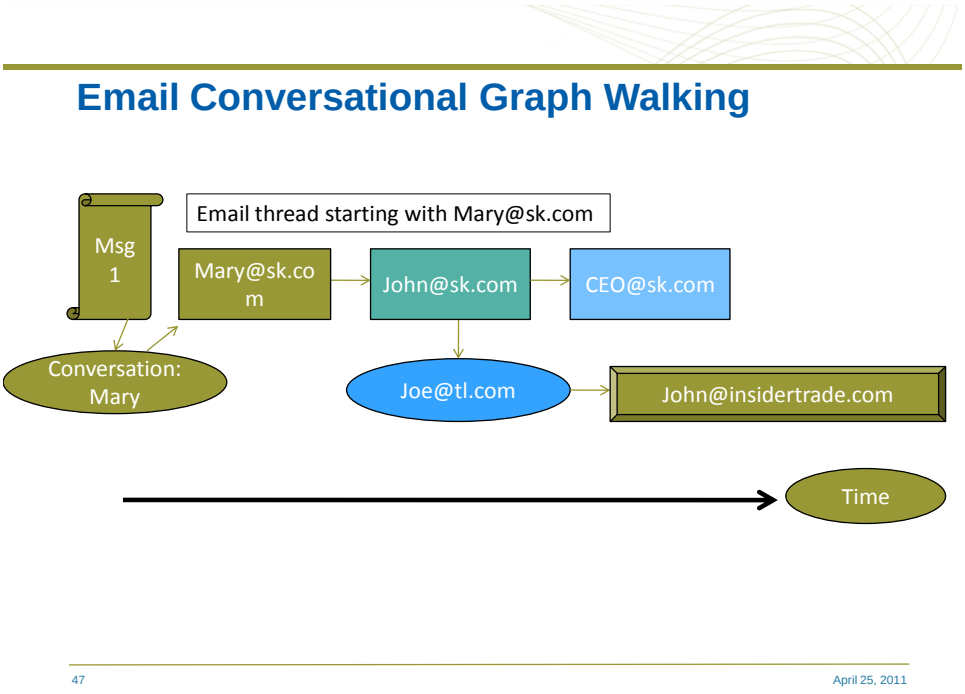
- Yielding a "difference" of one word between the two documents; after "little" and before "lamb". So the difference would be defined as an "insertion" between "little" and "lamb".

Mixing this kind of capability with the timeline that could be derived for when the edits occurred (by analyzing the Meta data that should be stored with the documents) both can be used to expose evidentiary facts about a case.

Email Conversational Analysis

Analyzing conversations within email and other message types is a very important part of legal discovery. A full email conversational analysis tool must include the ability to see what individuals sent messages on certain topics to others. In addition, it is important to have a tool that can display the full list of email "domains" that a person used in a given time period. This explains the main sites or companies contacted by a given individual over a specific period of time.

Figure Fifteen: Email Analysis



There are several approaches to email conversational analysis. The important aspect of this is allowing the correct attributes (both Meta data (header) information and content) to be included in the algorithm that constructs or follows the conversations.

Legal Discovery Processing Versus "Pure" Natural Language Processing

As we saw in the previous sections, there are a number of techniques that one can use to find conceptual relationships across document collections. In a desire to use the latest computer science

techniques to discover information users of legal review technology have turned to Natural Language Processing (NLP) and information analysis approaches that have been used in search engines and e-commerce applications.

Unfortunately legal review professionals have often turned to vendors who confuse general NLP techniques with sound legal discovery practice. The notion that a single mathematical technique will identify everything within a data set that is relevant to a case and help produce all relevant documents as a result is incomplete thinking. NLP techniques when applied correctly can be very helpful and powerful; but like any tool they can only be used in the correct circumstances. Also, at certain magnitudes of scale, the techniques break down, or experience limitations in their feasibility to produce relevant results. Over-fitting can be a problem that obfuscates results and makes the burden of computation a luxury for what the techniques provide in benefits to the review process.

This is why this paper started off with the full explanation of what a platform for legal discovery needs to contain. If the user understands that multiple operations need to be supported to find all aspects of the information relevant to a case (keyword data, Meta data, user-supplied Meta data, analytic associations) then NLP techniques can be one of those operations and the user will have great results. If the system a user selects relies on some specific NLP technique alone, the results it produces will not be complete enough for legal review purposes.

Data Preparation is Important to Obtaining Appropriate Results

We saw in previous sections that the way documents are prepared and represented within a legal discovery system is very important to obtaining good results. If data is not prepared correctly, certain techniques will break down (such as phrase searching performance). With analytics, stemming may reduce the dimensions an algorithm must analyze, but that may yield less specific results than one would envision.

In legal discovery, it can be very important to find documents that say: “by taking the following action; the party is ***choosing*** to violate the contract”. If the documents in a collection are prepared for “NLP” approaches, more documents than one really wants will be returned when looking for the phrase shown above or documents may be missed in the review phase. The near-duplicate mechanisms shown can find too many items that are not true “near-duplicates” if stemming is utilized. So one-step approaches must be carefully scrutinized if the most relevant results are to be obtained.

This can require extra human review and perhaps lead to human error during review. Many products prepare their collections of documents one way (to support both NLP and keyword search approaches). It is important to prepare documents specifically for the type of analysis (NLP or straight keyword-phrase searching) they will undergo. For legal discovery, it is important to prepare collections that can return results specific enough to save time in the initial collection reduction and the eventual legal review portions of the process.

The total platform approach (with the virtual index) lets one prepare data for the analytic operations that are important to each stage of a legal discovery process. This is possible because the virtual index

can represent the same data in multiple ways. Along with this, it is important to realize the benefits that analytics can provide.

Aspects of Analysis Algorithms for Legal Discovery

Another aspect of processing data for legal discovery and utilizing NLP techniques is that language characteristics of documents must be taken into account. Most NLP techniques use the statistical nature of data (via token frequency or occurrence) to derive some sort of model that describes data of certain types. If documents containing multi-lingual characteristics are combined with English-only documents, the predictive power of the model will decrease.

If language is not properly accounted for, the predictive power can become even less precise than it would be otherwise. Data preparation is very important to analytic performance in these types of systems. Legal review requires more precision than other applications so it is especially important to be precise with the preparation of data sets.

Benefits of Analytic Methods in Ediscovery

In an Ediscovery context, analytics are very important. They help the reviewer in several ways:

1. They can expose information about a collection of documents that is non-obvious but that can help one understand the meaning of the information they contain. This can help a lawyer understand what keywords would be relevant to a matter, and to select the ones that ultimately get used to discover information about a legal matter.
2. They can identify relationships in data that reveal what information was available to the parties to a lawsuit at certain points in time.
3. They can identify versions of documents and relate these to a timeline to make a reviewer aware of how knowledge related to a lawsuit or regulatory matter has evolved over time.
4. They can be used to find the documents that relate to known example documents within a collection. This helps a reviewer find all documents that are relevant and can also help a reviewer find other relevant concepts that may not have been in an initial keyword search list.

Problems with Analytic Procedures in Ediscovery

As stated above in the introduction to this section of the document, a major problem with analytic procedures in Ediscovery is that one technique is not appropriate in all circumstances. As vendors have tended to champion one technology for their analytics, they tend to promote the over-use of one technique that is available through the use of their particular technology. In their desire to find “the holy grail” or identify the “magic bullet” for legal review users often grab on to technology pushed forward from a certain vendor and then find that it is not the panacea that it was supposed to be.

Once they realize that this is an issue, some customers buy what they perceive as best of breed products. For legal discovery this has historically meant multiple ones; some for analytics and others for keyword search; perhaps a third or fourth for legal processing. Users typically try to use them separately. Outside of a single platform these technologies lose some of their value because loading data into and unloading data from various products introduces the chances of human and other error.

The introduction of one platform that can handle multiple analytic approaches is how one confronts the fact that there is no single analytic technique that masters all problems with electronic discovery.

Related to this issue of multiple products is that the products on the market do not run at the scale necessary to add value in even a medium sized legal matter. Because of this, analytic procedures are (practically) run after a data set has been reduced in size. This can be appropriate, but it can also reduce the useful scope and overall usefulness of the analytic technique in question. If some analytic techniques are run on a very large data set they can take an inordinately long time to run, making their value questionable. In addition, some techniques "break-down" after a certain scale and their results become less useful than they are at lower document counts.

An Ideal Platform Approach

To combat these issues with analytics, the correct platform with a scalable architecture and the appropriate "mix" of analytics is proposed as the answer. In the following sections a set of techniques that have been developed to ameliorate most of the issues with well-known analytic approaches will be shown.

The platform approach includes a "two-tier" ordering algorithm that first "sorts" data into related categories so that deeper analysis can be undertaken on groups of documents that belong together (at least in some sense). This helps the second-level algorithm run at appropriate scale and even avoid "bad choices" when sampling documents for running analysis that can identify conceptual information within documents. This is possible because of the grid architecture explained above and the correct mix of analytic techniques.

Analytic Techniques in Context of a Legal Discovery "Ideal Platform"

So given the assertion that no single analytic technique is adequate on its own to provide legal discovery analysis, this section discusses how a single platform using a combination of different analytic techniques could be valuable. In addition, it shows how a platform implementing several techniques allows the overall system to provide better results than if it had been implemented with one single analytic technique.

The ideal discovery platform:

1. Uses a specific and powerful initial unsupervised classification (clustering) technique to organize data into meaningful groups, and identifies key terms within the data groups to aid the human reviewer. Other analytic processes can take advantage of this first order classification of documents as appropriate
2. Uses a powerful multi-step algorithm and the grid architecture to organize data which has semantic similarity; conceptual cluster groups are formed after accounting for language differences in documents
3. Allows the user to select other analytic operations to run on the classification groups (folders) built in the first unsupervised classification step. This allows other analytic algorithms to be run at appropriate scale and with appropriate precision within the previously classified data folders

(LSA or PLSA for example) the benefit would be that the ideal platform could break the collection down and then allow PLSA or LSA to run at an appropriate scale if a judge ordered such an action

4. Allows the user to select documents from within the folders that have been created
    a. Using keyword search
    b. Using visual inspection of automatically applied document tags
    c. Via the unsupervised conceptual clustering techniques
5. Allows the user to select documents from folders and use them as example documents
    a. "Search by document" examples where the entire document is used as a "model" and compared to other documents
    b. Examples that can be used as "seed" examples for further supervised classification operations
6. Allows the user to tag and otherwise classify documents identified from the stage one classification or from separate search operations
7. Allows the user to identify predominant "language groups" within large collections of documents so that they can be addressed appropriately and cost effectively (translation, etc.)

Conceptual Classification in the Ideal Platform

As we learned in an earlier section of this document, this is an analytic technique that answers the question: "what is in my data"? It is designed to help a human reviewer see the key aspects of a large document collection without having to read all the documents individually and rank them. In some sense it also helps a reviewer deduce what the data "means". In the context of this discussion, it should be noted that the user of this functionality does not have any idea about what the data set contains and does not have to supply any example documents or "training sets".

Conceptual classification supports a number of uses within the ideal discovery product. These include:

1. Organizing the data into "folders" of related material so that a user can see what documents are semantically related; also it builds a set of statistically relevant terms that describe the topics in the documents
2. Presenting these folders so that search results can be "tracked back" to them. This allows a user to use keyword search and then select a document in the user interface and subsequently see how that document relates to other documents the unsupervised classification algorithm placed with the one found from keyword search. This is possible because the virtual index contains the document identifiers and the classification tags that show what related information exists for a given document.
3. Allows other "learning algorithms" to use the classification folders to identify where to "sample" documents for conceptually relevant information (explained below). This means that a first-order unsupervised classification algorithm orders the data so that other analytic processes can select documents for further levels of analysis from the most fruitful places in the document group. This allows higher-order language models (LSA, PLSA or n-gram analysis) to be run on them with a finer-grained knowledge of what the data set contains and to avoid sampling documents and adding their content to a model of the data that might make it less powerful or

predictive. This allows the system to identify the best examples of information where higher-level analysis can reveal more meaningful relationships within document content. Building models of similar documents from a previously unseen set of data is a powerful function of a system that contains analysis tools.

Unsupervised Conceptual Classification Explained

This technique solves the problems that were seen above with the single-technique approach (LSA/PLSA) where over-fitting can become an issue and specificity of results is lost. This technique:

1. Orders the data initially into folders of related material using a linearly interpolated statistical co-occurrence calculation which considers:
    a. Semantic relationships of absolute co-occurrence
    b. Language set occurrence and frequency
    c. This stage of the algorithm does NOT attempt to consider polysemy or synonymy relationships in documents; this is considered in the second stage of the algorithm
2. Performs a second-level conceptual "clustering" on the data where concepts are identified within the scope of the first-level "Clusterings". A latent generative technique is used to calculate the concepts that occur in the first-level cluster groups. This portion of the algorithm is where synonymy and polysemy are introduced to the analysis; "lists" of concepts are computed per each first-level or first-order cluster group; these may be left alone or "merged" depending on the results of the stage three of the algorithm
3. The "lists" of second-level conceptual cluster groups are compared; concepts from one folder are compared to those computed from another. If they are conceptually similar (in cross-entropy terms) they are combined into a "super-cluster". If they are not similar, the cluster groups are left separate and they represent different cluster groups within the product
4. The algorithm completes when all first-order clusters have been compared and all possible super-clusters have been formed

Algorithm Justification

This algorithm allows the data to be fairly well organized into cluster groups after the first-level organization. More importantly, it removes documents from clusters where they have nothing in common, such as documents primarily formed from foreign language (different character set) data. This is important because most latent semantic algorithms will consider information that can be irrelevant (on a language basis) thus obfuscating the results of the concept calculations. This also localizes the analysis of the conceptual computations. Secondly, the over-fitting problem is reduced because the latent concept calculations are undertaken on smaller groups of documents. Since conceptual relationships can exist across the first-level folder groups, the concept lists can be similar; denoting the information in two folders is conceptually related. The third step of the algorithm allows these similarities to be identified and the folders "merged" into a "super-folder" or "super-cluster" as appropriate. Therefore the result is a set of data that is conceptually organized without undue over-fitting and dilution of conceptual meaning.

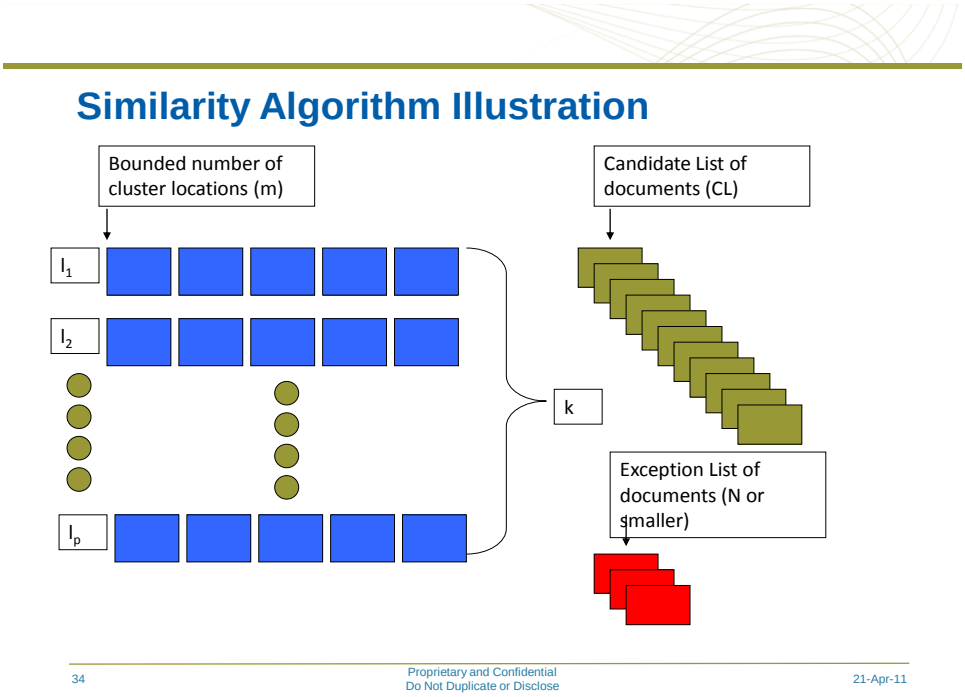
The trick to unsupervised learning or classification is in knowing where to start. The algorithm for unsupervised classification automatically finds the documents that belong together. The algorithm starts by electing a given document from the corpus as the "master" seed. All subsequent seeds are selected relative to this one. This saves vast amounts of processing time as the technique builds lists that belong together and "elects" the next list's seed automatically as part of the data ordering process.

Other algorithms randomly pick seeds and then try to fit data items to the best seed. Poor seed selection can lead to laborious optimization times and computational complexity. With the ideal platform's linear ordering algorithm, the seeds are selected as a natural course of selecting similar members of the data set for group membership with the current seed. There will naturally be a next seed of another list which will form (until all documents have been ordered).

First-Order Classification

This seed-selection happens during the first-order organization of the data. The aim of this part of the algorithm is to build "lists" of documents that belong together. These lists are presented to the users as "folders" that contain "similar" items. The system sets a default "length" of each list to 50 documents per list. The lists of documents may grow or shrink or disappear altogether (the list can "lose" all of its members in an optimization pass); initially the list is started with 50 members however. Please see Figure Sixteen for an illustration of the clustering technique.

Figure Sixteen: Unsupervised Classification



The algorithm starts with a random document; this becomes the first example document or "seed" to which other documents are compared. The documents are picked from the candidate documents in the collection being classified (candidate list; or every document in the corpus initially). There are "N" of these documents in the collection. The lists are of length "m" as shown in the illustration (as stated the

default value of m is 50); the number of lists is initially estimated at k=N/m. The first document is a seed and all other documents are compared to this document; the similarity calculation determines what documents are ordered into the initial list (l_1).

One key aspect of this technique is that initially all documents are available for selection for the initial list. Each subsequent list only selects documents that remain on the candidate list however. This is a "linear reduction" algorithm where the list selections take decreasing amounts of time as each list is built. There is a second optimization step to allow seeds that did not have a chance to select items that were put on preceding lists to select items that "belong" (are more similar to) them and their members.

Each document is compared to the initial seed. The similarity algorithm returns a value between 0 and 1; a document with exact similarity to a seed will have a value of 1, a document with no similarity (nothing in common) will have a value of zero. The candidate document with the highest similarity to the seed is chosen as the next list member in l_1. When the list has grown to "m" members the next document found to be most similar to the seed S_1 is chosen as the seed for the next list (l_2). The list l_2 is then built from the remaining documents in the candidate list. Note that there are N-m members of the candidate list after l_1 has been constructed. This causes (under ideal conditions) a set of lists, with members that are related to the seeds that represent each list, and with seeds that have something in common with one another.

Similarity Calculation

This similarity calculation takes into account how often tokens in one document occur in another and account for language type (documents that contain English and Chinese are "scored" differently than documents that contain only English text). This is a linearly interpolated similarity model involving both the distance calculation between data items and the fixed factors that denote language type. A document that would have a similarity score of "0.8" relative to its seed (which has only English text), based on its English text content alone, but that has a combination of English, Chinese and Russian text will have a score that is "lower" than 0.8 because the semantic similarity score will be reduced by the added attributes of the document having all three languages. This way the system can discern documents that have very similar semantics but that have different languages represented within them. The three language types are viewed as three independent statistical "events" (in addition to the language co-occurrence events of the tokens in the two documents). All events that occur within the document influence its overall probability of similarity with the seed document.

With some "language-blind" statistical scoring algorithms it is possible to have document scores which represent a lot of commonality in one language (English) and where the presence of Chinese text does not influence this much at all. If specific language types are not added into the calculation of similarity, documents with three different languages will appear to be as significantly similar to a seed as those which have only one language represented within their content.

Please see Figure Seventeen for an illustration of the first stage of the algorithm. Please see Figures Seventeen through Figure Twenty Two for other aspects of the algorithm.

Figure Seventeen: "Normal Operation" (Step One)

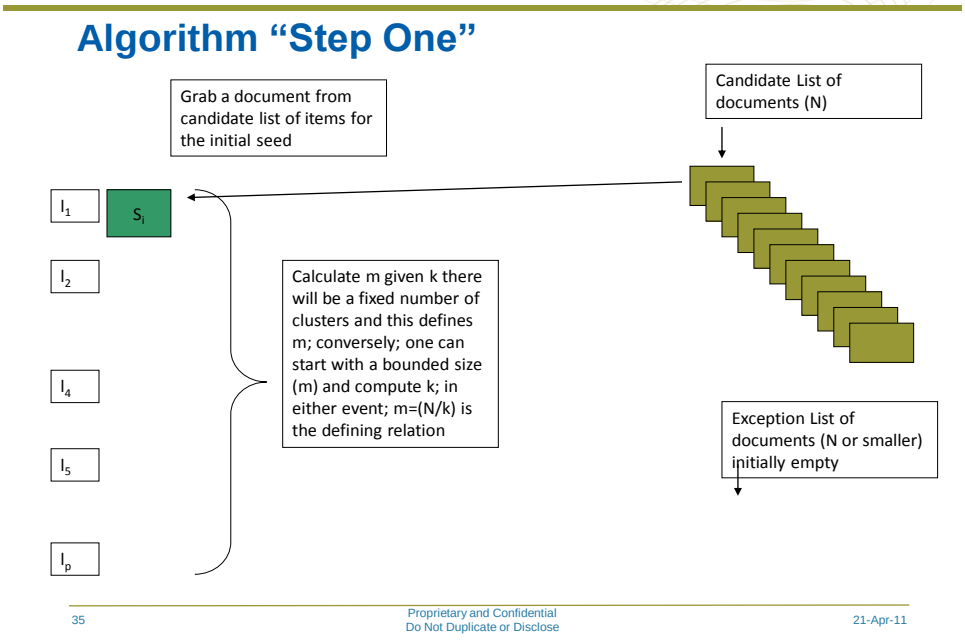


Figure Eighteen: Normal Operation

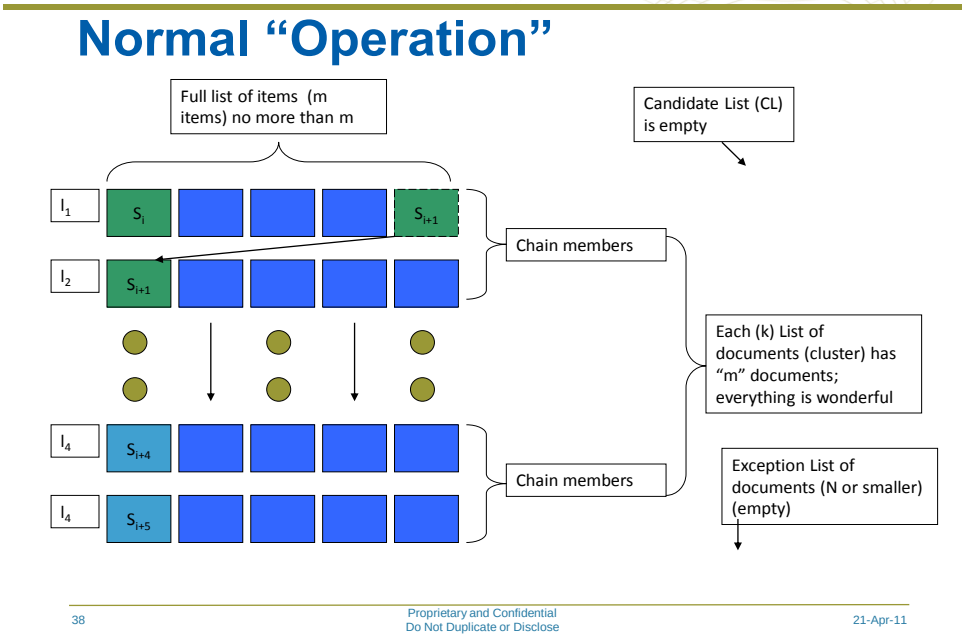


Figure Nineteen: List Formation

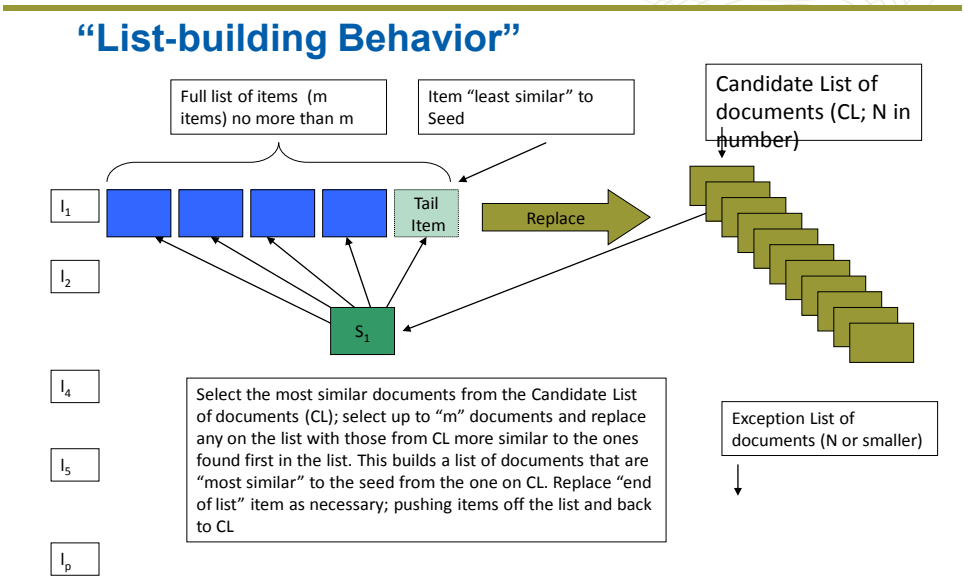
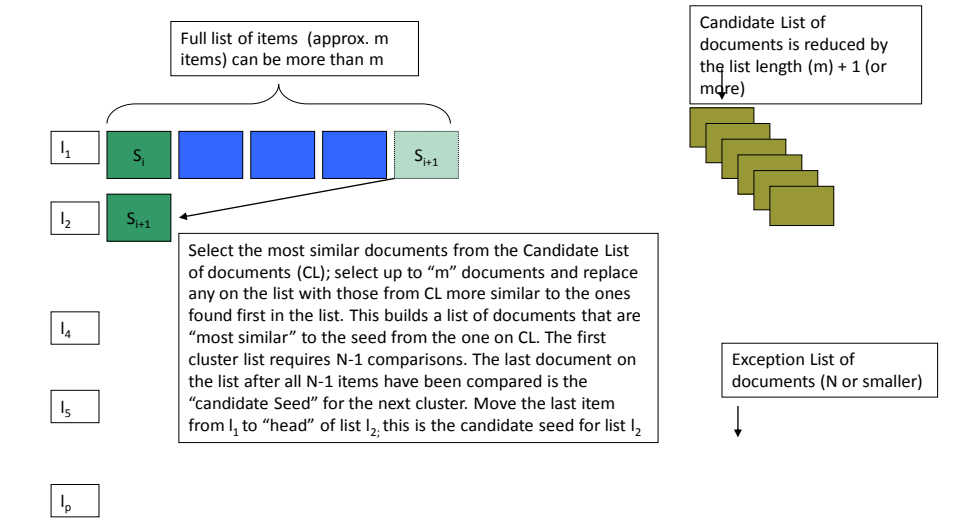


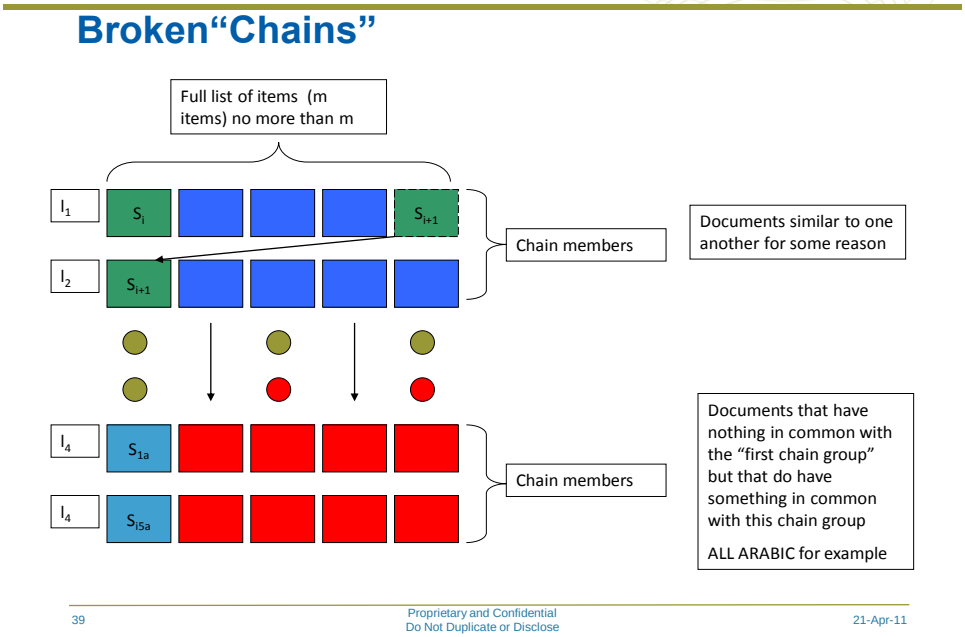
Figure Twenty: Linear Set Reduction



The lists which form under normal operation where there are documents with some similarity to a given seed for a given list are as shown in Figure Eighteen. The seed of each new list is related to the seed of a prior list and therefore has a transitive similarity relationship with prior seeds. In this respect each list that has a seed related to a prior seed forms a "chain" of similarity within the corpus. The interesting thing that occurs is when a seed cannot find any documents that are similar to it. This occurs when a chain of similarity "breaks" and in many cases, a "new chain" forms. This is when a seed cannot find a relationship in common with any remaining item on the candidate list. The document similarity of all remaining members of the candidate list, relative to the current seed is zero. This causes the chain to break and the seed selection process to begin again. Please see Figure Twenty One for an illustration of this behavior.

When a seed in a prior list does not find any items on the candidate list which is "similar" to it the algorithm selects an item from the candidate list as the next seed and the process starts over again. If items that are similar to this newly selected seed document exist on the candidate list, they are selected for membership in a new list (headed up by the newly selected seed) and the new list forms the head of a new chain.

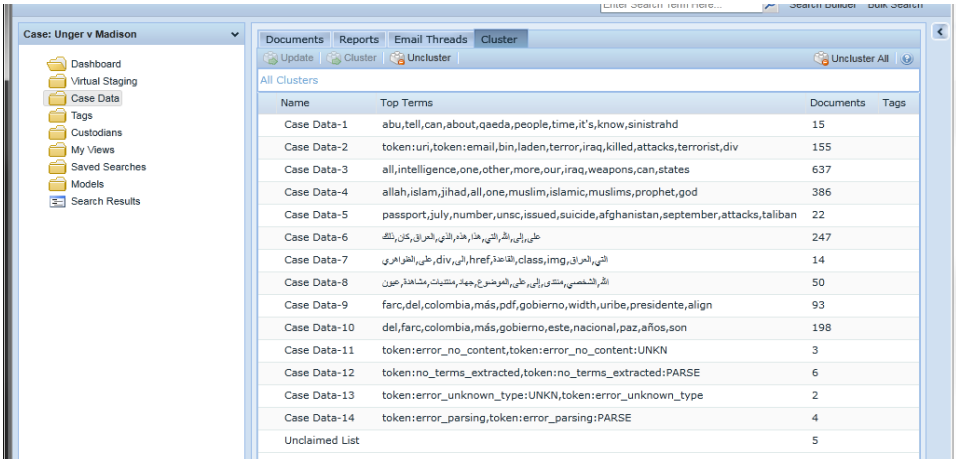
Figure Twenty One: Broken Chains



In Figure Twenty One it is shown that the documents in the first chain group have no commonality with documents in the second chain group. The documents in the second chain group do have some commonality among themselves however. The second chain group is formed by selecting a new seed at random from the candidate list when a seed from the first chain group finds no documents in the candidate list with which it has attributes in common. This phenomenon indicates a major change in the nature of the data set. This major change is often related to the document corpus having a set of

documents from a totally different language group than that represented in the first chain of lists and documents. This occurs when the language of the documents in a given chain group are English and the next chain group is Arabic for example. Figure Twenty Two is an actual screen shot from a product that shows a “cluster” of documents that are composed of Arabic text. This same behavior can occur within the same language group, but this is the most common reason that it occurs.

Figure Twenty Two: Chain Behavior Displayed in Classification Groups



In this example the folders labeled: “Case-Data 6”, “Case-Data 7” and “Case Data 8” contain Arabic text documents. This occurred because the textual similarity of the documents had little to do with English and were very much in common because of the Arabic text they contain. The similarity algorithm put the Arabic documents together because the product evaluates each “token” of text as it is interpreted in Arabic. The frequency of Arabic tokens in the documents compared with “English seeds” showed no similarity with a given English document seed. An Arabic “chain” formed and attracted Arabic documents to these particular folders. Similar assignments happened in these data for Spanish and French documents.

Second-Order Classification

It was explained above, but with the benefit of the illustration it is clear that the conceptual calculations are undertaken on the folders consecutively. They benefit from the fact that the overall calculation has been broken into groups that bear some relationship to a seed document that leads the cluster. Even if conceptual similarity spans two clusters, the third and final stage of the algorithm will “re-arrange” the cluster membership to order the documents conceptually. Computing concepts on each individual cluster from the first stage of the algorithm reduces the number of documents in the calculation and thus over-fitting.

The technique used in the ideal platform at this stage is reviewing conditional probabilities with prior statistical distributions. It is drawing an initial “guess” of how the terms in the documents are distributed statistically by gathering information about the document classifications found in the first-order classification step. It computes the likelihood of certain terms being “topics” within certain documents

and within the overall collection of documents. It orders topics and documents so that they can be regarded as “topic labels” for the documents that are contained within the folders.

Third-Order Classification

This stage of the algorithm will re-order any documents into the final folder groups according to the conceptual similarity of the concept lists computed in stage two of the algorithm. Documents which “belong” to a “super cluster” are merged to be with the documents that are most similar to the concept list computed for some number of second stage clusters. Concept lists for each second stage cluster are compared and if their members are similar, the documents forming the lists are merged into a final super cluster; otherwise the documents are left in the cluster they inhabit. This allows conceptually similar folders to be merged together; the documents comprising folders with similar concept lists will be re-organized into a super-cluster. The documents the folders represent “belong together” so the folders are “merged”.

First-Order Classification Importance

As stated previously, when documents contain different languages, the algorithms that compute the labels for document groups can lose precision. These algorithms look at the probabilities of certain terms occurring with other terms and when multiple languages are involved their results can become skewed. The first-order classification algorithm puts the documents with similar first-order language characteristics together, which aids the performance of the second-order topic generation algorithm. The two algorithms together are more powerful than either one is together. This also helps the third-order classification algorithm as the folders that have similar characteristics tend to be “near” one another in the folder list. Even if they are not, the concept clustering algorithm will find the conceptually related information that “goes together” but merging is often possible early on in the “walking” of the folder concept lists because of the first-order classification operation.

Second-Order and Final Classification Importance

With this technique, the topics that are latent are generated by the algorithm running on the folders of pre-ordered documents themselves. This provides the benefits of LSA or PLSA on the pre-ordered sets of data but with much less computation (by taking advantage of the classification from the first pass of the algorithm). Without the first-order technique, more computation would be required to arrive at the optimal generation of the topics. This would place a computational burden on the system unnecessarily. The first-order classification of the documents assists the second algorithm and makes further analysis much more “clear” as well. When the final check is done on the semantic label lists of the folders in the collection, they allow for the documents that belong together conceptually to be re-clustered as necessary.

Multi-lingual Documents

It is important to note that the algorithm still handles multi-language documents, and the concept generation algorithm can find cross-correlations of terms in multiple languages. Terms that occur in a document containing English and Arabic text will have conceptual lists that contain both Arabic and English members. The first-order grouping of primarily Arabic or English documents together will still

allow for single language correlations to predominate but will not prohibit the generation of concepts from documents containing both Arabic and English text.

Value of Classification

The value of classification like this is that a human reviewer can quickly identify document groups that may be of interest. The "top reasons" or "top terms" that a folder contains (the predominant terms in the documents it contains) is shown at the top of the folder in the screen shot contained in Figure Twenty Two. The user of the product can determine if the documents are of any interest to him/her quickly by reading the labels on each folder. Further, the reviewer does not have to open and read documents that may be Arabic or Chinese (unless they want to read them). This folder based ordering of documents allows a reviewer to avoid obviously irrelevant information such as documents in a language that is of no interest to them. More importantly however, this pre-ordering first-stage classification technique makes higher-order analysis of the data in the folders accurate and predictable and more computationally efficient. The end-stage classification yields strong language semantics for members of final classification groups.

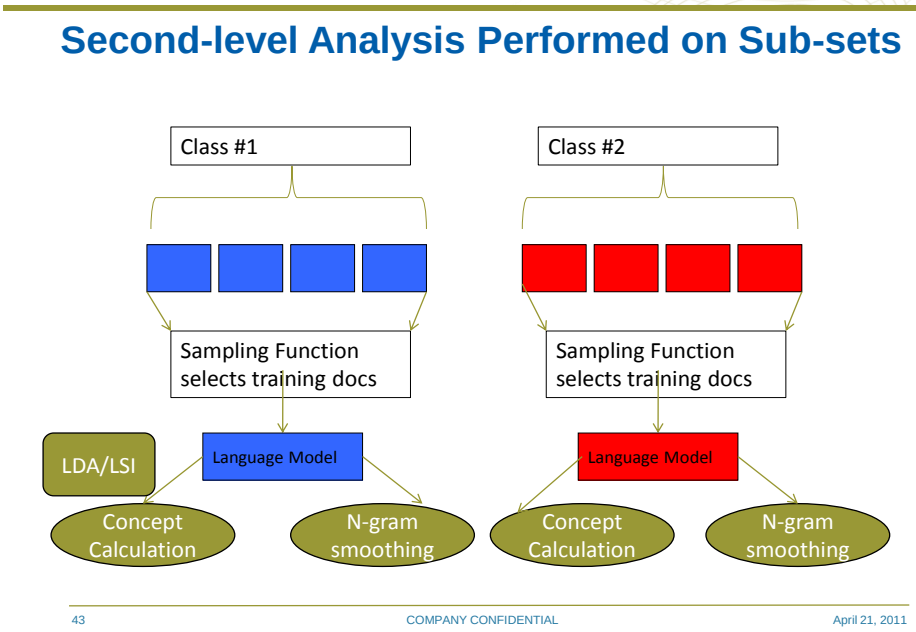
Other Benefits of Document Classification

Other benefits are that any document found with a keyword search can be related back to the classification groups. Since the documents in these classification folders are related to their seed, they are very likely to be related to one another. This allows a reviewer to see what other documents are similar to a given document found via keyword search where the reviewer knows that the reference document contains at least agreed upon key terms. The conceptual organization past this step allows for the final clusters to include semantically conceptual clustering relationships that can be related back to keyword searches. The ideal system also allows one to use the pre-ordering classification for other analysis techniques such as PLSA or LSA. Either would benefit from operating on pre-ordered data that is smaller than the entire collection.

N-gram or Other Statistical Learning Models

For building language specific tools on top of the base classification engine, the pre-ordering technique is especially useful. After the classification algorithm has ordered a collection, higher-order language models can be built from samples within the larger collection. This may be beneficial for building n-gram models for language specific functions like part of Speech (POS) tagging. Other tools that could benefit from this would be learning models that support functions such as sentence completion for search query operations. In these cases, knowing that a cluster group contains primarily Arabic textual information would allow the n-gram model to select samples from an appropriate set of documents. This can be important if one is building a model to handle specific functions such as these. The first-order algorithm will "mark" the analytic Meta data for a certain cluster group to show that it represents a predominant language. For a POS tagger, this would be important as many of these are highly sensitive to the input model data and training them with appropriate samples is important. It would be counter-productive to train an English language POS tagger with German training data for example. The first-order algorithm allows one to select documents from the appropriate places within the larger collection of documents for specific purposes. See Figure Twenty-Three for an illustration of this behavior.

Figure Twenty Three: Selecting from Pre-Classified Data for Higher-Order Models



How Pre-Ordering Can Make Other Techniques Better

This first-order classification capability can reduce the amount of documents that any second order algorithm has to consider. This can help other less-efficient algorithms run more effectively. As with the concept calculation example (above) it may be desirable to run LSA or other tools on data that the platform has processed. By utilizing the folder-building classification algorithm within the platform, the large population of documents for a given case could be reduced to more manageable sized increments that an algorithm like LSA can handle.

If opposing counsel were to insist on running LSA or PLSA or some other tool from a select vendor on a data set, the ideal platform could order and organize smaller folders of data that LSA or PLSA could handle. This technique of pre-ordering the data will generate smaller sized related folders that these other techniques could process at their more limited scale. The reduced size of the data set would help focus the results of LSA because it would have fewer documents and find fewer concepts to fit into the "buckets". Therefore the platform could help reduce the LSA over-fitting issue. This would help PLSA in this regard as well.

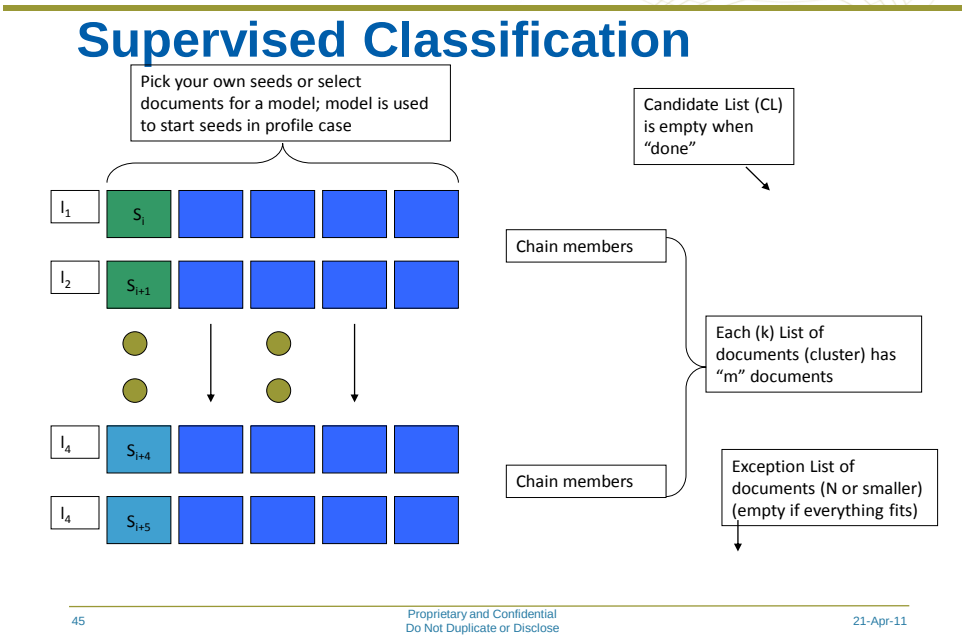
As mentioned in other sections of this document, LSA was envisioned and works best on collections of documents that are "smaller" than those that are routinely found during current legal discovery matters. In [4] it is stated by the authors of the LSA algorithm that they envision "reasonable sized" data sets of "five thousand or so" documents. In today's cases, it is routine to see 100,000 to 200,000 or more documents (millions are not uncommon).

With the combined platform approach a technique like LSA could be run on the documents from the most likely "clusters" from the ideal platform so that the computation would be tractable. If a judge felt comfortable with LSA as a technique due to prior experience with the algorithm he or she could see better results by reducing the amount of documents that the algorithm has to address at any one time. The use of the ideal platform would benefit a legal review team by making a previously used product implementing LSA more effective at what it does.

Supervised Classification

Supervised Classification requires the user to provide some example data to which the system can compare unclassified documents. If a user has some documents that they know belong to a certain classification group, a system can compare documents to the examples and build folders of documents that lie within a certain "distance" from the seeds (in terms of similarity).

Figure Twenty Four: Supervised (Example Based) Classification



If the system has pre-ordered a lot of the data (using unsupervised techniques as before), then finding examples is simplified for the user. They have some idea where to look for examples that they can use to classify documents that will enter the case as new documents. Secondly, search results can be related to document clusters that exist, then examples of the "strongest most similar" documents to the search results can be located within a cluster folder, and then the supervised classification technique can identify other documents that belong with pre-selected "seeds". Again, documents added to a case can be classified with examples using the supervised technique shown above. This continuous classification

of documents can help reviewers find the most relevant documents rapidly with more or less automated means.

Email Conversational Analysis

Email conversational analysis is an important aspect of any legal review platform. Seeing what conversations transpired between parties is important. This was discussed previously. With the ideal platform approach of providing classification along with the email threading, these two techniques can be used simultaneously to identify documents that are in a conversation thread, and that have similar documents which may exist outside that thread. The existence of documents that are similar to those in a thread will lead to the identification of email addresses that perhaps were outside the custodian list but that should be included. Having the conversation analysis and the classification capability all within one platform makes this "analytic cross-check" capability possible.

Near-Duplicate or Version Analysis

The version analysis mentioned above, combined with supervised clustering and Meta data search can identify what documents were edited at certain times and by whom. Using near duplicate analysis can allow the system to "tag" all members of a "near-dupe group" within a collection of documents (auto-tagging of analytic Meta data). Using supervised clustering with a known seed (from the near dupe group) a user can identify other versions within the collection of documents comprising a case. Using the Meta data attributes to identify owners and import locations of documents that have been collected exposes information about who owned or copied version of files at any time during the life cycle of the case. This is a very powerful attribute of a platform that handles these combined sets of analytic processes at large scale.

Summary

This paper attempted to display the value behind a comprehensive platform that handles various levels of indexing for Meta data content and analytic structures. It exposed new concepts behind analysis and storage of these constructs that implement high-speed indexing and analysis of data items within the context of legal discovery. Further, it explored and discussed several aspects of large scale legal discovery processing and analysis and how the correct architecture combined with indexing and search capabilities can make legal discovery effective and productive relative to current single product approaches.

The "ideal platform" approach (as it was called) presented both architecture and a set of capabilities that remove risk of error from legal discovery projects. Examples of how this combination would reduce cost and risk of error in legal discovery engagements were presented.

Finally, analytic approaches that are available from electronic discovery products today, how they work, where they are effective and where they are not effective were presented. These were compared and contrasted with one another and were discussed in relation to the ideal platform approach. The ideal platform and its ability to pre-order and classify data at great scale, and then perform generative concept and label generation to identify the "meaning" of content and assign it to "folders" of

documents within large cases was discussed. It was shown how the platform approach of pre-classifying data and using a hierarchical model of classification algorithms could aid other products and techniques such as those that utilize LSA and PLSA.

References

[1] Wikipedia description of Latent Semantic Analysis: http://en.wikipedia.org/wiki/Latent_semantic_analysis

[2] Wikipedia description of Probabilistic Latent Semantic Analysis: http://en.wikipedia.org/wiki/PLSA

[3] Wikipedia description of Bayesian Classifiers: http://en.wikipedia.org/wiki/Bayesian_classification

[4] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman (1990). "Indexing by Latent Semantic Analysis"

[5] Wikipedia page on Polysemy: http://en.wikipedia.org/wiki/Polysemy

[6] Wikipedia page on Support Vector Machines: http://en.wikipedia.org/wiki/Support_vector_machine

[7] Wikipedia page on Vector Space Model: http://en.wikipedia.org/wiki/Vector_space_model

[8] Wikipedia page on SMART system: http://en.wikipedia.org/wiki/SMART_Information_Retrieval_System

[9] Web Page of Martin Porter: http://tartarus.org/~martin/PorterStemmer/

[10] Wiki entry on mathematical treatment of SVD: http://en.wikipedia.org/wiki/Singular_value_decomposition

[11] Wiki entry on LSA/LSI: http://en.wikipedia.org/wiki/Latent_semantic_indexing

[12] Adam Thomo Blog: mailto:thomo@cs.uvic.ca] Blog entry for LSI example

[13] ^ Thomas Hofmann, *Probabilistic Latent Semantic Indexing*, Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99), 1999

bit-x-bit computer forensic investigation and e-discovery consulting services

The Frick Building
437 Grant Street, Suite 1501
Pittsburgh, Pennsylvania 15219
Tel: 412-325-4033
www.bit-x-bit.com

BIT-X-BIT, LLC

DESI IV POSITION PAPER

APRIL 18, 2011

This Position Paper describes the interests of bit-x-bit, LLC ("bit-x-bit") in the Fourth DESI Workshop.

bit-x-bit is computer forensics and e-discovery consulting firm with offices in the Frick Building in downtown Pittsburgh, Pennsylvania. We work with companies, law firms, and other clients in Pittsburgh and throughout the country, providing e-discovery services such as collection of electronically stored information ("ESI") for litigation, ESI processing (de-duplification, de-NISTing, indexing), key word searching, testing of key words, hosted review tools, ESI document preparation and production. We also perform computer forensic investigations for use in civil and criminal cases. We are exclusively endorsed by the Allegheny County Bar Association as the provider of e-discovery and computer forensic services to its more than 6,500 members. Recently, we planned and prepared the materials, assembled a panel, and presented the four-hour Orientation Program for the E-discovery Special Masters ("EDSM") admitted to the EDSM Program in the U.S. District Court for the Western District of Pennsylvania.

We are interested in participating in the Workshop because we work with a wide variety of lawyers, from very small firms to very large firms. Virtually all of our e-discovery clients use key word searching as a means of identifying potentially relevant and responsive documents for litigation. bit-x-bit assists its client-litigants in selecting and refining key words in order to maximize the effectiveness of search terms in locating and recalling potentially relevant documents. We are interested in state of the art techniques to make key words and related search technology as precise and efficient as possible, so as to minimize review time and money spent on the review of irrelevant documents. As "predictive coding" techniques and software are developed, we are interested in considering such tools for our use, and the use of our clients, in setting standards for the case and project management techniques necessary to use effectively such software, and in the research, testing and data which supports the use of such tools and the defensibility of such tools.

Respectfully submitted,

Susan A. Ardisson, Esq., CEO

W. Scott Ardisson, President

Joseph Decker, Esq., Vice President and General Counsel

A Perfect Storm for Pessimism: Converging Technologies, Cost and Standardization

Topics: Information Retrieval, Data Mining, Machine Learning (Supervised and Unsupervised), eDiscovery, Information Governance, Information Management, Risk Assessment, Computational Intelligence
Author: Cody Bennett
Company: TCDI - http://www.tcdi.com

If some vendors' proclamations are to be believed, realizations of next generation self-learning technologies geared towards text retrieval, language understanding and real time risk assessments are being fulfilled. But the knowledge experts in charge of assisting these platforms need to be aware of exiguous claims. If standardization is going to occur on a matrix of such complex systems, they need to occur *on reality, not hype*.

The amount of information will grow vastly while storage costs become subdued increasing the need for computational technologies to offset the very large costs associated with knowledge workers. This paradigm shift signals a mandatory call for smarter information systems, both automated and semi-automated. But imperfections in technology systems (arguably lower than human mistakes) require critical focus on workflows, modularity and auditing. And although linear systems have improved efficiencies through the use of virtualization, they still do not approach a lateral learning mechanism[1]. While they have begun to break into multi-tenant super-computing capacity, software on these systems is still statistical and rules-based, hardly approaching anything "thinking" – encompassing decades-old algorithms stigmatized by "Artificial Intelligence".

Further, the cost prohibitive business model reliant upon a single technology becomes a landmine. Technology in the space of Information Retrieval and Machine Learning are moving targets, and formal standardizations may be quickly outmoded. While there are attempts to use previous and ongoing research alongside industrial search studies performed to classify and understand the limitations of each search model[2], use of hybridization and the underlying platforms / architectures facilitating multiple types of search techniques should be a target for which Information Management systems strive. For eDiscovery, vendors should be prepared to harness multiple search capabilities as courtrooms over time mold what is accepted as "standard". Focusing on a single methodology when coupled with automated systems hampers recall – IBM's Watson and Intelligence organizations prove that hybridized multimodal search and brute force NLP based directed probabilistic query expansion are interesting because of combinations in Information Retrieval, Data Mining and Machine Learning. How do you standardize upon the algorithms entrenched in systems that are constantly in flux? Do only systems with little or no entropy deserve standardization?

Use of multimodal search is becoming fashionably effective in tandem with automation. Machine Learning methods utilizing hybrid approaches to maximize historically divergent search paradigms are capable of producing multiple viewpoints based on different algorithms, maximizing return on implementations such as predictive coding, active "tripwire" systems and next-generation risk assessment. In eDiscovery, multiple modeling viewpoints can help augment linguistic spread of features necessary to defensibly identify varying degrees of responsiveness. An example would be the improvement for the eDiscovery process using active learning when conducting initial discovery and query expansion / extrapolation in the beginning phases of Request for Production[3].

With both Information Retrieval and Machine Learning, transparency in the methods and a heavy breakdown of the algorithms used will be required. This transparency assists Information Governance, defensible methods for legal, and quality assurance for knowledge workers. This prognostication may be similar to the inevitability of eDiscovery certification in bar exams. While it may not be necessary for legal to understand the full complexities of the underlying search technology or automated algorithm, it should be required to ascertain and request certifiable tests meeting standardized thresholds on retrieval software and learning systems especially in comparison with human counterparts. These standards not only directly affect industry and researches in academia, but legal teams who may view such technology as foreign. Legal in the realm of Information Governance will become the centrality for delivering the dos and don'ts of information in the corporation, in partnership with the CIO / IT, and possibly as oversight.

More robust search algorithms and sophistication in automated apparatuses allow more document discovery to be performed. While it could be argued by legacy eDiscovery review shops that such systems displace

1 See Lateral learning and the differentiators of Artificial and Computational Intelligence
2 NIST TREC, Precision, Recall, F-measure, ROC
3 http://trec.nist.gov/pubs/trec19/papers/bennett.cody.LEGAL.rev.pdf

workers, the resulting outcome will be more time for their expertise to focus on larger data sets and cases. The technology tools also allow new forms of discovery. During litigation, if both counsels are using automated methods, expect different forms of data mining and statistical modeling to look for fringe data; Information Governance becomes critically important because signposts to documents that were not produced may become evident. It also puts the onus on the automated systems. Though, even while precision, speed and capacity may massively increase, the chance of sanctions should increase less dynamically dependant upon the unknowns of the output. In review, knowing that automated coding will always make the same calls if the parameters and data remain the same may be comforting. But the hive instinct of a group of humans making judgments on the fly is tempered when replaced by the efficiency. Are vendors willing to champion their products against comparisons of human reviewers in standardized sessions? Are they willing to "open up the hood" for transparency?

Along with the many previous buzzwords, possibly the biggest is "Cloud". Information Management, Cloud and semi / automated eDiscovery provide historically high potential for low cost, immediate, real-time view into the information cycle. Which means, not only will businesses entertain cloud services, but because of lower cost, less worry about infrastructure, and touted uptime, they will be able to search and store more information as they adhere to rules for retention and preservation. Whether a public or private cloud or some hybrid, this growth of searchable data will necessitate further automation of processes in Information Governance and solidification of the underlying framework – policies, procedures and standards beyond search of information.

The standardization for Clouds may be best lead by the Government and related agencies. Cost of Government is under heavy scrutiny and current endeavors are occurring to facilitate the movement of Government into the Cloud. Cloud infrastructure, believing the hype, will structurally allow the computing capacity needed for today's brute force systems and experimental Computational Intelligence *et al*[4]. This intriguing ability to perform massive calculations per second with elasticity is a lowly feature compared to the perceived cost savings which currently drives the interest for mid to large sized entities; public clouds like Microsoft, Amazon and Salesforce.com currently among the most popular. Although, for eDiscovery, the cost of demanding and actually acquiring documents from geographically disparate locations may produce a haven for sanctions. More ominously, if mission critical systems become cloud based, could critical infrastructure (industry, state, and government) become even more exposed[5]?

This architecture triangulation (Cloud + [Enterprise] Information Retrieval + Machine Learning) is either a Nirvāṇa or the Perfect Storm. Whatever viewpoint, the criticality is security. Providing a one stop shop for data leaks and loss, hack attacks, whistle blowing and thievery across geographically massive data sets of multitudes of business verticals combined with hybridized, highly effective automated systems designed to quickly gather precise information with very little input at the lowest possible cost is one CIO's wish and one Information Manager's nightmare[6]. Next generation systems will need to work hand in hand with sophisticated intrusion detection, new demands for data security and regulators across state and international boundaries – and hope for costs' sake, that's enough. Standardized security for different types of clouds was bluntly an afterthought to cost savings.

Finally, technology growth and acceptance while cyclic is probably more spiral[7]; it takes multiple iterations to conquer very complicated issues and for such iterations to stabilize. Standardizing Artificial, Computational and Hybrid Intelligence Systems is no different. The processes underneath these umbrella terms will require multiple standardization iterations to flesh out the bleeding edge into leading edge. It is possible that the entropy of such systems is so high that standardization is just not feasible. Where standardization *can* occur in the triangular contexts described above, expect it to follow similar structure as RFCs from the Internet Engineering Task Force[8]. Though, this will likely require heavy concessions and the potential unwillingness from industry on interoperability and transparency.

---

4 Pattern analysis, Neural Nets

5 A next generation Stuxnet, for example…

6 Not to mention, lawyers holding their breath in the background…

7 This type of cyclical information gain when graphed appears similar to Fibonacci (2D) and Lorentz (3D) spirals.

8 This makes sense due to the fact that data access and search has been spinning wildly into the foray of Internet dependence.

WILLIAMS MULLEN
Where Every Client is a Partner®

May 2011

EDIG: E-Discovery & Information Governance

The Williams Mullen Edge

Why Document Review is Broken

BY BENNETT B. BORDEN, MONICA MCCARROLL, MARK CORDOVER & SAM STRICKLAND[1]

The review of documents for responsiveness and privilege is widely perceived as the most expensive aspect of conducting litigation in the information age. Over the last several years, we have focused on determining why that is and how to fix it. We have found there are several factors that drive the costs of document review, all of which can be addressed with significant results. In this article, we move beyond costs and get to the real heart of the matter: document review is a "necessary evil" in the service of litigation, but its true value is rarely understood or realized in modern litigation.

It was not always so. When the Federal Rules of Civil Procedure were first promulgated in 1938, they established a framework from the common law with respect to which discovery took place. But there was no fundamental change in how one conducted discovery of the comparatively few paper documents that comprised the evidence in most civil cases. There was no Facebook or even email at the time. Only later, when the sheer number of paper documents grew to a point where litigators needed help to get through them, and only later still when the electronic creation of documents became possible and then ubiquitous, did the "problem" of information inflation convert document review into a separate aspect of litigation, and one that accounted for a significant portion of the cost of litigation.

There are three primary factors that drive the cost of document review: the volume of documents to be reviewed, the quality of the documents, and the review process itself. The volume of documents to be reviewed will vary from case to case, but can be reduced significantly by experienced counsel who understands the sources of potentially relevant documents and how to target them narrowly. This requires the technological ability to navigate computer systems and data repositories as well as the legal ability to obtain agreement with opposing counsel, the court or the regulator to establish proportional, targeted, iterative discovery protocols that meet the needs of the case. Because of the important work of The Sedona Conference® and other similar organizations, these techniques are better understood, if not always widely practiced.[2]

At some point, however, a corpus of documents will be identified that requires careful analysis, and how that "review" is conducted is largely an issue of combining skillful technique with powerful technology. In order to take advantage of all of the benefits this technology can provide, the format of the documents, the data and metadata, must be of sufficient quality. When the format of production is "dirty" (i.e., inconsistent, incomplete, etc.), you face a situation of "garbage in/garbage out." For several reasons, "garbage" in this sense no longer suffices.

As we discuss more fully below, the most advanced technology we have found uses all of the aspects of data and metadata to improve the efficiency (and thus reduce the cost) of the review process – and more. This means that the ESI must be obtained, whether from the client for its own review or from opposing counsel for the review of the opposing party's documents, with sufficient metadata in sufficiently structured form to capitalize on the power of the technology. This requires counsel with technological and legal know-how to obtain ESI in the proper format. Many negotiated ESI protocols have become long and complex, but they rarely include sufficiently detailed requirements concerning the format of documents, including sufficiently clean data and metadata such that the most powerful technologies can be properly leveraged. Without this, a great deal of efficiency is sacrificed.

Williams Mullen EDIG Team

Bennett B. Borden
Co-Chair
804.420.6563
bborden@williamsmullen.com

Monica McCarroll
Co-Chair
804.420.6444
mmccarroll@williamsmullen.com

Stephen E. Anthony
757.629.0631
santhony@williamsmullen.com

Jonathan R. Bumgarner
919.981.4070
jbumgarner@williamsmullen.com

W. Michael Holm
703.760.5225
mholm@williamsmullen.com

William R. Poynter
757.473.5334
wpoynter@williamsmullen.com

Brian C. Vick
919.981.4023
bvick@williamsmullen.com

Lauren M. Wheeling
804.420.6590
lwheeling@williamsmullen.com

Ada K. Wilson
919.325.4870
awilson@williamsmullen.com

Once a corpus of documents has been identified and obtained in the proper format, the document review commences. This is where we have found the greatest inefficiency, and this is the primary area in which the most significant gains are possible. Our analysis of the typical review process leads us to conclude that the process is broken. By this we mean that, typically, document review is terribly inefficient and has been divorced from its primary purpose, to marshal the facts specific to a matter to prove a party's claims or defenses and to lead to the just, speedy and inexpensive resolution of the matter. This disheartening conclusion led us to question whether document review could be completed efficiently and effectively within days or even hours so that a party could almost immediately know its position with respect to any claim or defense. That kind of document review could become an integral part of the overall litigation as well as the primary driver of its resolution.

But document review has become an end unto itself, largely divorced from the rest of litigation. The typical review is structured so that either contract attorneys or low-level associates conduct a first level review, coding documents as responsive, non-responsive or privileged. Sometimes the responsive documents are further divided and coded into a few categories. But this sub-dividing is usually very basic and provides only the most general outline as to the subjects of the documents. Typically, a second level review is conducted by more senior associates to derive and organize the most important facts. Thus, every responsive document is reviewed at least twice, and usually several more times as the second level reviewers distill facts from the documents to organize them into case themes or deposition outlines that are finally presented to the decision makers (usually partners).

This typical tiered review process is inherently inefficient and requires a great deal of time and effort. The most pressing question that arises in the beginning of a matter, "what happened?", prompts the answer, "We'll tell you in two (or three or six) months." This multiplicitous review process leads to lost information in transfer, lost time, and the attendant increase in cost. The three standard categories (responsive, non-responsive and privileged) result in oversimplification because not all responsive documents are equally responsive. Add to inefficiency, then, simple misinformation. Is this avoidable?

Document review became separated from the litigation process because of the increase in the volume of potentially relevant documents. With thousands or even millions of documents to review, law firms or clients typically threw bodies at the problem, hiring armies of contract attorneys to slog through the documents one by one in an inefficient linear process. The goal was simply to get through the corpus to meet production deadlines. But, if the whole point of document review is to discover, understand, marshal and present facts about what happened and why, then it is the facts derived from document review that drive the resolution of the matter. Thus, the entire discovery process should be tailored to this fundamental purpose. Part of this, as we have noted, must be accomplished through experienced counsel who understands what the case is about, what facts are needed, and how to narrowly and proportionally get at them. The other key is to derive facts from the reviewed documents as quickly and efficiently as possible, and transfer the knowledge distilled from those facts to the decision makers in the most effective and efficient way. In short, document review should be returned to its rightful place as fact development in the service of litigation.

In October 2010, we released an article entitled: *The Demise of Linear Review*, wherein we discussed the use of advanced review tools to create a more efficient review process.[3] There, we showed that by using one of these advanced tools, and employing a non-linear, topic-driven review approach, we were able to get through several reviews between four and six times faster than would be the case with less advanced tools using a typical linear review approach. Since then, we have focused on perfecting both the review application and our review processes. Our results follow below.

The application used in the reviews described in *The Demise of Linear Review* was created by IT.com and is called IT-Discovery (ITD). The non-linear, topic-driven reviews were conducted by a team of attorneys led by Sam Strickland, who has since created a company called Strickland e-Review (SER), and the reviews were overseen by the Williams Mullen Electronic Discovery and Information Governance Section. The ITD application uses advanced machine learning technology to assist our SER reviewers in finding responsive documents quickly and efficiently, as we showed in *The Demise of Linear Review*. But we wanted to show not only that our topic-driven review process was faster, but also that it was qualitatively better than a typical linear review. Here we move into the area of whether humans doing linear review are in fact better than humans using computer-assisted techniques - not only for cost reduction but for improving the quality of results. We tested this and concluded that humans doing linear review produce significantly inferior results compared to computer-assisted review.

To prove this, we obtained a corpus of 20,933 documents from an actual litigation matter. This corpus had been identified from a larger corpus using search terms. The documents were divided into batches and farmed out to attorneys who reviewed them in a linear process. That review took about 180 hours at a rate of about 116 documents per hour. The typical rate of review is about 50 documents per hour, so even this review was more efficient than is typical. Our analysis showed that this was because the corpus was identified by fairly targeted search terms, so the documents were more likely to be responsive. Also, the document requests were very broad, and there were no sub-codes applied to the responsive documents. Both of these factors led to a more efficient linear review.

We then loaded the same 20,933 documents into the ITD application and reviewed them using our topic-driven processes with SER. This review took 18.5 hours at a rate of 1,131 documents per hour, almost ten times faster than the linear review. Obviously it is impossible for a reviewer to have seen every document at that rate of review, so we must question whether this method is defensible. To answer that question, it is important to distinguish between reviewing a document and reading it.

Reviewing a document for responsiveness means, at the most fundamental level, that the document is recognized, through whatever means, as responsive or not. But this does not mean that the document has to be read by a human reviewer, if its responsiveness can otherwise be determined with certainty. The ITD application uses advanced machine learning technology to group documents into topics based upon content, metadata and the "social aspects" of the documents (who authored them, among whom were they distributed and so forth), as well as the more traditional co-occurrence of various tokens and forms of matrix reductions that constitute modern machine learning techniques to data mine text. Because of the granularity and cohesiveness of the topics created by ITD, the reviewers were able to make coding decisions on groups of documents. But more interestingly, these unsupervised-learning-derived topics aid in intelligent groupings of all sorts, so that a reviewer can "recognize" with certainty a large number of documents as a cohesive group. One can then code them uniformly.

Does this mean that some documents were not "reviewed" in the sense that a reviewer actually viewed them and made an individual decision regarding their responsiveness? No. To understand this by analogy, think of identifying in a corpus all Google news alerts by the sender, "Google alert," from almost any advanced search page in almost any review or ECA platform, in a context where none of these documents could be responsive. Every document was looked at as a group, in this case a group determined by the sender, and was coded as "non-responsive." This technique is perfectly defensible and is done in nearly every review today. What we can do is extend this technique much deeper to apply it to all sorts of such groups and voilá: a non-linear review on steroids.

Isn't there some risk, however, that if every document isn't read, privileged information is missed inadvertently and thus produced? Not if the non-linear review is conducted properly. Privileged communications occur only between or among specific individuals. Work product only can be produced by certain individuals. Part of skillful fact development is knowing who these individuals are and how to identify their data. The same holds true for trade secrets, personal information, or other sensitive data. The key is to craft non-linear review strategies that not only identify responsive information, but also protect sensitive and privileged information.

We showed in our non-linear review that the SER reviewers, using the advanced technology of the ITD application, coded 20,933 documents in $1/10^{th}$ of the time that it took the linear reviewers to do so. The question then becomes, how accurate were those coding decisions? To answer this question, we solicited the assistance of Maura R. Grossman and Gordon V. Cormack, Coordinators of 2010 TREC Legal Track, an international, interdisciplinary research effort aimed at objectively modeling the e-discovery review process to evaluate the efficacy of various search methodologies, sponsored by the National Institute of Standards and Technology. With their input, we designed a comparative analysis of the results of both the linear and non-linear reviews.

First, we compared the coding of the two reviews and identified 2,235 instances where the coding of the documents conflicted between the two reviews. Those documents were then examined by a topic authority to determine what the correct coding should have been, without knowing how the documents were originally coded. Results: The topic authority agreed with the ITD/SER review coding 2,195 times out of 2,235, or 98.2% of the time.

Not only was the ITD/SER review ten times faster, it resulted in the correct coding decision 99.8% of the time. In nearly every instance where there was a dispute between the "read every document" approach of the linear review and our computer-assisted non-linear review, the non-linear review won out. Could this just be coincidence? Could it be that the SER reviewers are just smarter than the "traditional" reviewers? Or perhaps, as we believe, is the fundamental approach of human linear review using the most common review applications of today simply worse? The latter position has been well documented by Maura R. Grossman and Gordon V. Cormack, among others.[4]

The implication of this specific review, as well as those discussed in *The Demise of Linear Review*, is that with our ITD/SER review process we can get through a corpus of documents faster, cheaper and more accurately than with traditional linear review models. But, as we have noted, document review is not an end unto itself. Its purpose is to help identify, marshal and present facts that lead to the resolution of a matter. The following is a real-world example of how our better review process resulted in the resolution of a matter. We should point out that case results depend upon a variety of factors unique to each case, and that case results do not guarantee or predict a similar result in any future case.

We represented a client who was being sued by a former employee in a whistleblower *qui tam* action. The client was a government contractor who, because of the False Claims Act allegations in the complaint, faced a bet-the-company situation. As soon as the complaint was unsealed, we identified about 60 custodians and placed them on litigation hold, along with appropriate non-custodial data sources. We then identified about 10 key custodians and focused on their data first. Time was of the essence because this case was on the Eastern District of Virginia's "Rocket Docket." We loaded the data related to the key custodians into the ITD platform and SER began its review before discovery was issued. We gained efficiency through the advanced technology of the ITD platform. We also gained efficiency by eliminating the need to review documents more than once to distill facts from them. Our review process includes capturing enough information about a document when it is first reviewed so that its facts are evident through the organizing and outlining features in ITD. This eliminated the need for a typical second- and even third-level review.

Within four days, the SER reviewers could answer "what happened?". Soon thereafter, the nine reviewers completed the review of about 675,500 documents at a rate of 185 documents per hour. More importantly, within a very short time we knew precisely

what the client's position was with respect to the claims made and had marshaled the facts in such a way as to use them in our negotiations with the opposing party, all before formal document requests had been served.

Knowing our position, we approached opposing counsel and began negotiating a settlement. We made a voluntary production of about 12,500 documents that laid out the parties' positions, and walked opposing counsel through the documents, laying out all the facts. We were able to settle the case. All of this occurred after the production of only a small fraction of all the documents, without a single deposition taken, and at a small fraction of the cost that we had budgeted to take the case through trial.

This real-world example demonstrates the true power of "document review" when understood and executed properly. Fundamentally, nearly every litigation matter comes down to the questions of "what happened?" and "why?". In this information age, the answers to those questions almost invariably reside in a company's ESI, where its employees' actions and decisions are evidenced and by which they are effectuated. The key to finding those answers is knowing how to narrowly target the necessary facts within the ESI. You then can use those facts to drive the resolution of the litigation. This requires the ability to reasonably and proportionally limit discovery to those sources of ESI most likely to contain key facts and the technological know-how to efficiently distill the key facts out of the vast volume of ESI.

The typical linear document review process is broken. It no longer fulfills its key purpose: to identify, marshal and present the facts needed to resolve a matter. Its failure is legacy to the nature of how it came into being as the volume of documents became overwhelming. We believe we have found the right combination of technique and technology to return the process to its roots, resolving litigation.

For more information about this topic, please contact the author or any member of the Williams Mullen E-Discovery Team.

---

[1] Bennett B. Borden and Monica McCarroll are Chairs of Williams Mullen's Electronic Discovery and Information Governance Section. Mark Cordover is CEO of IT.com. Sam Strickland is President of Strickland e-Review.

[2] See, Bennett B. Borden, Monica McCarroll, Brian C. Vick & Lauren M. Wheeling, *Four Years Later: How the 2006 Amendments to the Federal Rules Have Reshaped the E-Discovery Landscape and are Revitalizing the Civil Justice System*, XVII RICH. J.L. & TECH. 10 (2011), http://jolt.richmond.edu/v17i3/article10.pdf.

[3] See, *The Demise of Linear Review*, October 2010, http://www.williamsmullen.com/the-demise-of-linear-review-10-01-2010/

[4] Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11 (2011), http://jolt.richmond.edu/v17i3/article11.pdf.

Editorial inquiries should be directed to Bennett B. Borden at 804.420.6563, bborden@williamsmullen.com or Monica McCarroll at 804.420.6444, mmccarroll@williamsmullen.com.

WILLIAMS MULLEN
Where Every Client is a Partner®

LITIGATION SERVICES

Planning for Variation and E-Discovery Costs

By Macyl A. Burke

President Eisenhower was fond of the quote, “In preparing for a battle I have found plans are useless, but planning is indispensable”. The same logic, creating a foundation and framework, can be applied effectually to the complex world of e-discovery where every case brings its own uniqueness and quirks.

That said, there are some check points that should be examined in triangulating the factors of cost, risk, and time that loom large in the world of complex litigation. Discovery is the most expensive piece in legal spend, with review as the most expensive element in discovery. Discovery and review are estimated at over 80% of the cost by some sources. It is therefore logical to be thoughtful about the discovery process in general and review in particular.

Cost Calculations

We suggest the following points be analyzed and considered in the decision making process of the legal discovery cost:

- ✓ HOW DO THEY CHARGE? Examine the basis for your Economic Model: Look for cost models that are document or gigabyte based. Explore alternative cost models that are fixed and transparent. Integrated costing models that combine both the cost of the law firm and the vendor on a fixed price basis can be advantageous. The critical path is to achieve the lowest all-in cost which includes the supervision of the law firm and their certification of the process. The line item costs taken in and of itself are not the critical path.

- ✓ WHAT DO THEY MEASURE? Know your enemy: You should know your blended hourly rate. The blended rate is the cost of the contract attorneys combined with the cost of the law firms to supervise and certify the review. This is usually the single largest cost in the discovery process and largest available cost saving opportunity in most projects. If you are not paying for the review by the document or gigabyte in an integrated cost model be sure to understand the blended rate.

- ✓ WHAT TOOLS DO THEY HAVE? Practice MTCO: “Measure Twice; Cut Once” is a best practice. There are numerous methods of measurement that can be applied to your project to cut cost and improve results. Just a few are blended rate, sampling, review rates, error rates, richness of the population, recall, precision, page counts, document counts, pages per

document, etc. that can be used to analyze your project. A vendor should have a tool box of metrics to explain and explore the best way forward. These metrics can be instrumental in reducing the size of the population to be reviewed in a reasonable good faith way and result in large direct savings in terms of all-in cost. Good metrics help foster innovation and are a catalyst to change.

- ✓ WHAT DO THEY REPORT? Use sampling over inspection: By all means necessary, avoid inspection as a quality assurance practice by the law firm in the review stage of discovery. Sampling is materially less expensive and produces a far superior result. Sampling will reduce the cost of your blended rate geometrically and provide more current and better results as to quality and how well the knowledge transfer has taken place. It allows for quick course correction and continuous improvement. Inspection is expensive rework that is not necessary.

- ✓ CAN THEY DO IT FOR LESS? Quality processes should lower cost: Understand the quality control and quality assurance processes being employed by the vendor and law firm. If they do not lower cost they are not quality applications. They are cost increases. The lowered costs should be measurable and produce the desired results.

- ✓ CAN THEY DO MORE? Select an integrated provider: The largest savings comes from reducing the discovery population to be reviewed as much as possible in a reasonable and good faith manner and reviewing it in a high quality low cost application. An integrated provider who offers processing, culling, hosting and first cut review is in the best position to achieve these efficiencies. Using an integrated provider allows for the growth and development of a strong partnership with the participating law firm.

- ✓ HOW DO I LEARN MORE? Take advantage of third party information: Use proven quality processes from other sources. The Sedona Conference Paper Achieving Quality in the e-Discovery Process is an excellent source. It references numerous other sources which are rich in information. Ralph Losey's blog is highly useful with a diverse set of contributors.

- ✓ Keep up to date on current developments: E-discovery is a fluid and dynamic field. The TREC 2009 Legal Track Interactive Task offers new insights into technology assisted review. Additionally, proportionality and cooperation are emerging as important factors in the discovery process.

- ✓ Construct an Audit Trail and Flow Chart: Document carefully the processes and results across the whole of the EDRM. You want a record that your process was reasonable, in good faith, and proportional. Be sure to document ongoing changes. Even your best planning will need to be adapted to emerging realities that could not have been anticipated or known.

Risk Evaluation

Variation in complexity, scale, cost, and risk are present in any system or process. Good quality is about reducing variation to acceptable and predictable levels which are confirmed by metrics. This is particularly true in the discovery process of litigation that represents the bulk of the legal spend. In general there has been some price compression on discovery activities but there is still a large amount of cost variation in the discovery process. The standard is that the discovery burden must be reasonable, in good faith, and proportional. By that standard, using measurement and understanding, the variation insures compliance with the requirement.

A few specific examples can show the scope of the problem that variation presents in the discovery process.

In an actual case, we audited one firm using contract reviewers at a cost of $53 an hour for 1st Tier review with law firm attorneys doing supervision and 2nd Tier review at $300 plus an hour which achieved an all-in cost of $7 a document. In the same case, we found another firm using associates for 1st Tier review at a cost of $250 an hour and 2nd Tier review and supervision for $300 plus an hour for an all-in cost of $20 a document. These were documents being reviewed in the same case with the same issues. The only difference was the cost and the process. In neither case were the results measured.

Evaluation: There is significant variation between the two firms and the costs do not include processing, hosting, or production. We participate in an alliance with a law firm that would offer an all-in price (collection, processing, hosting, 1st and 2nd Tier review, certification, and production) or total cost of approximately $1.63 a document.

In another example, we audited a 1st Tier review at cost of $0.05 per page. There was not an hourly rate or per document rate involved. There were approximately 50,000 documents with a page count of 7,400,000 pages all of which were billed by the page. This works out to an average document page count of around 150 pages per document.

Evaluation: On a per page basis at $0.05 per page that turns out to be around $370,000. A more common per document charge in the range of $0.70 to $1.10 a document works out between $35,000 and $55,000.

Looking at various per gigabyte all-in cost numbers, we find the variation is enormous. The average all-in cost (collection, processing, hosting, 1st Tier and 2nd Tier review, certification, and production would be approximately $70,000 a gigabyte plus or minus $35,000. In our view, with a good integrated process the cost should be approximately $24,000 a gigabyte plus or minus $2,000 for all-in cost.

All of the pricing examples above offer large variations in outcomes. In planning, it is a good practice to measure from two different approaches and compare how close the results are to one another. We would recommend you examine several economic models for each project. Ask, how are we measuring results? What are the metrics, what is the cost per document, per gigabyte, etc? The difference between pricing by the page or document can be extreme.

Time Frame

The examples show the order of magnitude, variation, and the potential savings available in the cost of discovery. Eisenhower was also fond of the statement, “What is urgent is seldom important, and what is important is seldom urgent”. In the hair on fire world of high stakes litigation this is not the conventional wisdom. However, the spiraling cost of discovery fueled by ever increasing volumes of ESI (electronically stored information) should give us cause to pause and take a hard look at process, variation, and measurement. Planning can be derailed or incomplete by the drama of law and press of events. The temptation to short cut the planning activity should be avoided even if the urgency is great. Planning is too important to be co-opted by urgency.

If a vendor offers a magic plan in the e-discovery maze, be wary. Only planning can prepare for variations, cause awareness of alternatives, help discover pitfalls, and refine our goals. The given check points are general concepts that can be expanded to more granular applications. The approach should be emergent, based on circumstance and need. It is basically a read and react scenario using concepts that may or may not be appropriate in a given circumstance. The suggestions are not new or radical. We are offering them as touch points to reduce costs, lower risk, and improve time. By no means are they a panacea or magic bullet to spiraling legal costs. We do suggest they are reasonable and good faith questions that should be asked and answered.

April 2011

Semantic Search in E-Discovery

Research on the application of text mining and information retrieval for fact finding in regulatory investigations

David van Dijk, Hans Henseler
Amsterdam University of Applied Sciences
CREATE-IT Applied Research
Amsterdam, the Netherlands
d.van.dijk@hva.nl, j.henseler@hva.nl

Maarten de Rijke
University of Amsterdam
Intelligent Systems Lab Amsterdam
Amsterdam, the Netherlands
derijke@uva.nl

***Abstract*— For forensic accountants and lawyers, E-discovery is essential to support findings in order to prosecute organizations that are in violation with US, EU or national regulations. For instance, the EU aims to reduce all forms of corruption at every level, in all EU countries and institutions and even outside the EU. It also aims to prevent fraud by setting up EU anti fraud offices and actively investigates and prosecutes violations of competition regulations. This position paper proposes to address the application of intelligent language processing to the field of e-discovery to improve the quality of review and discovery. The focus will be on semantic search, combining data-driven search technology with explicit structured knowledge through the extraction of aspects, topics, entities, events and relationships from unstructured information based on email messages and postings on discussion forum.**

Keywords: E-Discovery, Semantic Search, Information Retrieval, Entity Extraction, Fact Extraction, EDRM

I. Introduction

Since the ICT revolution took off around 50 years ago the storage of digital data has grown exponentially and is expected to double every 18 months [16]. Digital data became of crucial importance for the management of organizations. This data also turned out to be of significant value within the justice system. Nowadays digital forensic evidence is increasingly being used in court. The Socha-Gelbmann Report from 2006 shows a usage of this kind of evidence in 60% of the court cases [31].

The process of retrieving and securing digital forensic evidence is called electronic data discovery (E-Discovery). The E-Discovery Reference Model [8] gives an overview of the steps in the e-discovery process. The retrieval of information from large amount of digital data is an important part of this process. Currently this step still involves a large amount of manual work done by experts, e.g. a number of lawyers searching for evidence in all e-mails of a company which may include millions of documents [30]. This makes the retrieval of digital forensic evidence a very expensive and inefficient endeavor [24].

Digital data in E-Discovery processes can be either structured or unstructured. Structured data is typically stored in a relational database and unstructured data in text documents, emails or multimedia files. Corporate Counsel [6] indicates that at least 50% of the material of contemporary electronic discovery environment is in the form of e-mail or forum and collaboration platforms. Finding evidence in unstructured information is difficult, particularly when one does not exactly know what exactly to look for.

The need for better search tools and methods within the area is reflected in the rapid growth of the E-Discovery market [32,10], as well as in the growing research interest [34,15,29]. This paper positions the research that is carried out through joined work from CREATE-IT Applied Research at the Amsterdam University of Applied Sciences [7] and the Intelligent Systems Lab Amsterdam at the University of Amsterdam [17]. It focuses on the application of text mining and information retrieval to E-Discovery problems.

II. Text Mining and Information Retrieval

Information retrieval (IR) can be defined as the application of computer technology to acquire, organize, store, retrieve and distribute information [19]. Manning defines IR as finding material (usually documents) of unstructured nature (usually text) from large collections (usually stored on computers) that provides in an information need [23]. Text mining (TM), also called text analytics, is used to extract information from data through identification and exploration of interesting patterns [9]. In TM, the emphasis lies on recognizing patterns. TM and IR have a considerable overlap, and both make use of knowledge from fields such as Machine Learning, Natural Language Processing and Computational Linguistics.

Both TM and IR provide techniques useful in finding digital forensic evidence in large amounts of unstructured data in an automated way. The techniques can be used for example to extract entities, uncover aspects of and relationships between entities, and discover events related to these entities. The extracted information can be used as metadata to provide additional guidance in the processing and review steps in E-

Discovery. Without such guidance, plain full-text search in large volumes of data becomes useless without proper relevance ranking. Metadata can be used to support interactive drill down search that is more suited for discovering new facts.

Furthermore, information about entities and aspects makes it possible to retrieve facts about a person as to what kind of position he currently holds, what positions he has previously had and what is important about him. Information about relationships can be used to identify persons closely connected with each other, but also to identify what persons are strongly connected to specific locations or (trans)actions. And events related to the entity can help one to extract temporal patterns.

III. Applications

The above techniques can be useful in many areas, both within and outside the domain of E-Discovery. Opportunities can be found in the areas of fraud, crime detection, sentiment mining (e.g., marketing), business intelligence, compliance, bankruptcies and, as one of the largest areas, e-discovery [27,12]. Large regulatory compliance investigations in the areas of anti-corruption and anti-trust offer excellent opportunities for text mining and information retrieval. Known techniques can be optimized and further developed to extract facts related to corruption and competition and to identify privileged and private information that should be excluded from the investigation.

For the detection of competition law infringements one can look at how prices develop [4]. For finding corruption one could search for suspicious patterns in transactions between entities, e.g., clients and business partners. In determining confidential data one can think of social security numbers, correspondence between client and attorney, medical records, confidential business information, etc. But often it is not clear beforehand what is sought, and therefore techniques are of interest that make the information accessible and provide insights so that a user can easily interact with it.

The entities and relations retrieved by the aforementioned techniques can be made accessible to the user in various ways. Either as additional metadata to documents to be combined with full-text search or as relational data in a separate system which can process questions in natural language (Question Answering System). The former gives a list of documents in response, the second can answer in natural language.

IV. Objective

Our research will focus on review and in particular on the search process. Generic search technology is not the answer. It has its focus on high precision results, where the top-ranked elements are of prime importance, whereas in forensic analysis and reconstruction all relevant traces should be found. In e-discovery, both recall and precision must be simultaneously optimized [26]. As a consequence, in e-discovery, the search process is typically iterative: queries are refined through multiple interactions with a search engine after inspection of intermediate results [5].

Analysts often formulate fairly specific theories about the documents that would be relevant and they express those criteria in terms of more-or-less specific hypotheses about who communicated what to whom, where, and, to the extent possible, why [2]. Representing, either implicitly or explicitly, knowledge associated with analysts' relevance hypotheses so that an automated system can use it, is of primary importance in addressing the key issues in e-discovery of how to identify relevant material [14]. Our research is aimed at providing analysts with more expressive tools for formulating exactly what they are looking for.

In particular, our research questions are as follows:

RQ1: At least 50% of the material in today's e-discovery environment is in the form of e-mail or forum and collaboration platforms [6]. How can the context (such as thread structure or the participant's history) of email messages and forum postings be captured and effectively used for culling entire sets of messages and postings (as they do not answer the question posed)?

RQ2: How can the diversity of issues that relate to the question posed be captured in a data-driven manner and presented to analysts so as to enable them to focus on specific aspects of the question?

RQ3: Social networks, graphs representing probable interactions and relations among a group of people, can enable analysts to infer which individuals most likely communicated information or had knowledge relevant to a query [28,13]. How can we effectively extract entities from e-mail messages and forum postings to automatically generate networks that help analysts identify key individuals?

RQ4: How can we semi-automatically identify the principal issues around the question posed? Creating an "information map" in the form of a domain-specific and context-specific lexicon will help improve the effectiveness of the iterative nature of the e-discovery process [36].

Based on typical user needs encountered in E-Discovery best practices, these research questions are situated at the interface of information retrieval and language technology. Answering them requires a combination of theoretical work (mainly algorithm development), experimental work (aimed at assessing the effectiveness of the algorithms developed) and applications (implementations of the algorithms will be released as open source).

V. Innovation

In recent years the field of information retrieval has diversified, bringing new challenges beyond the traditional

text-based search problem. Among these new paradigms is the field of semantic search, in which structured knowledge is used as a complement to text retrieval [25]. We intend to start a research project which pursues semantic search along two subprojects:

Subproject 1: integrating structured knowledge (discussion structure, topical structure as well as entities and relations) into information retrieval models;

Subproject 2: extracting structured knowledge from user generated content: entities, relations and lexical information.

We have requested funding for two PhD students, one for each of the two subprojects. Subproject 1 will primarily address RQ1 and RQ2. Subproject 2 will focus on RQ3 and RQ4.

Work on RQ1 will start from earlier work at ISLA [35] and extend the models there with ranking principles based on thread structure and (language) models of the experience of participants in email exchanges and collaborative discussions.

Work on RQ2 will take the query-specific diversity ranking method of [11], adapt them to (noisy) social media and complement them with labels to make the aspects identified interpretable for human consumption and usable for iterative query formulation.

Work on RQ3 will focus on normalization, anchoring entities and relations to real-world counterparts as captured in structured information sources. This has proven to be a surprisingly hard problem [20]. So far, mostly rule-based approaches have been used in this setting; the project will break down the problem in a cascade of more fine-grained steps, some of which will be dealt with in a data- driven manner, and some in a rule-based step, following the methodology laid down in [1].

Finally, in work on RQ4, principal issues in result sets of documents will be identified through semi-automatic lexicon creation based on bootstrapping, using the initial queries as seeds [21].

For all the questions described above we plan to conduct experiments in which we will implement our newly designed techniques and evaluate them by measuring commonly used metrics like precision and recall. By experimenting with different designs and evaluating them we expect to reach the desired level of quality expected from these techniques. Evaluation will take place by participating in benchmarking events like TREC [33], CLEF [3] and INEX [18] and by cooperating with business organizations within the stated areas.

As the aim of the TREC Legal Track [34] is to evaluate search tools and methods as they are used in the context of e-discovery, participating in this track seems to be an attractive way to start of our project. We will join the 2011 track with our first implementation for which we will use the Lemur Language Modeling Toolkit [22], complemented with implementations of the lessons learned at ISLA in the work referenced above. The track will provide us with workable data, focus, a deadline and it will provide us with a first evaluation of our work.

VI. Relevance for quality in E-Discovery

This research derives its relevance for quality in E-Discovery from three factors:

First, the research connects with the present (and growing) need of trained E-Discovery practitioners. Both national and international regulators and prosecutors are facing a large increase in the amount of digital information that needs to be processed as part of their investigations.

Second, the research is relevant for legal processes, as it directly addresses evidential search. The proceedings of their investigations impact in-house and outside legal counsel who are acting on behalf of companies that are under investigation. Intelligent language processing techniques can be a solution to effectively discover relevant information and to filter legal privileged information at the same time. This is not only a Dutch problem but also extends to international cases with US en EU regulators.

Third, the research will result in (open source) web services that can be exploited in E-Discovery settings. For testing and development purposes, open sources and/or existing data sets are available.

These factors and the active involvement of E-Discovery practitioners will be realized through their involvement in use case development, data selection and evaluation. We expect that this combination will increase the effectiveness and the quality of E-Discovery while information volumes will continue to explode.

REFERENCES

[1] Ahn, D., van Rantwijk, J., de Rijke, M. (2007) A Cascaded Machine Learning Approach to Interpreting Temporal Expressions. In: Proceedings NAACL-HLT 2007.

[2] Ashley K.D., Bridewell, W. (2010) Emerging AI & Law approaches to automating analysis and retrieval of electronically stored information in discovery proceedings. Artificial Intelligence and Law, 18(4):311-320.

[3] CLEF: The Cross-Language Evaluation Forum, http://www.clef-campaign.org/

[4] Connor, John M., (2004). How high do cartels raise prices? Implications for reform ofthe antitrust sentencing guidelines, American Antitrust Institute, Working Paper.

[5] Conrad J., (2010). E-Discovery revisited: the need for artificial intelligence beyond information retrieval. Artificial Intelligence and Law, 18(4): 321-345.

[6] Corporate Counsel, (2006). The American Bar Association (ABA), section of litigation, committee on Corporate Counsel. http://www.abanet.org/litigation/committees/corporate/

[7] CREATE-IT applied research - Onderzoek/lectoren, http://www.create-it.hva.nl/content/create-it-applied-research/onderzoeksprogrammas/

[8] EDRM: Electronic Discovery Reference model, http://www.edrm.net/

[9] Feldman, R., and Sanger, J. (2006). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.

[10] Gartner, (2009).MarketScope for E-Discovery Software Product Vendors report 2009, http://www.gartner.com/DisplayDocument?id=1262421

[11] He, J., Meij, E., de Rijke, M. (2011) Result Diversification Based on Query-Specific Cluster Ranking. Journal of the American Society for Information Science and Technology, to appear.

[12] Henseler, J.,(2010A). Openbare les E-Discovery: Op zoek naar de digitale waarheid. Amsterdam University of Applied Sciences.

[13] Henseler, J.,(2010B). Network-based filtering for large email collections in E-Discovery. Journal Artificial Intelligence and Law,Volume 18, Number 4, p.413-430

[14] Hogan C, Bauer R, Brassil D (2010) Automation of legal sensemaking in e-discovery. In: Artificial Intelligence and Law, 18(4):321-345

[15] ICAIL 2011: The Thirteenth International Conference on Artificial Intelligence and Law, http://www.law.pitt.edu/events/2011/06/icail-2011-the-thirteenth-international-conference-on-artificial-intelligence-and-law/

[16] IDC 2007: Research report on the Information Explosion.

[17] ISLA: Intelligent Systems Lab Amsterdam, http://isla.science.uva.nl/

[18] INEX: Initiative for the Evaluation of XML Retrieval, http://www.inex.otago.ac.nz/

[19] Jackson P., Moulinier I., (2002). Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization. Amsterdam: John Benjamins Publishing Company.

[20] Jijkoun, V., Khalid, M., Marx, M. de Rijke, M. (2008) Named Entity Normalization in User Generated Content. In: Proceedings of the second workshop on Analytics for noisy unstructured text data (AND 2008), pages 23-30, ACM.

[21] Jijkoun, V., de Rijke, M., Weerkamp, W. (2010) Generating Focused Topic-specific Sentiment Lexicons. In: 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010).

[22] LEMUR: Language Modeling Toolkit http://www.lemurproject.org/lemur/

[23] Manning, C. D., Raghavan, Prabhakar, and Schütze, Hinrich (2008). Introduction to Information Retrieval. Cambridge University Press.

[24] Oard, D.W., Baron, J.R., Hedin, B., Lewis, D.D., Tomlinson, S.,(2010). Evaluation of information retrieval for E-discovery. Journal Artificial Intelligence and Law,Volume 18, Number 4, p.347-386

[25] Pound, J., Mika, P., Zaragoza, H. (2010) Ad-hoc object retrieval in the web of data. In WWW 2010, pp 771-780.

[26] Rosenfeld L., Morville, P. (2002) Information architecture for the World Wide Web, 2nd edn. O'Reilly Media, Sebastopol.

[27] Scholtes, J. C., (2009). Text mining: de volgende stap in zoektechnologie. Inauguratie. Maastricht University

[28] Schwartz, M.F., Wood, D.C.M. (1993) Discovering shared interests using graph analysis. Communications of the ACM 36:78–89

[29] Sedona: The Sedona conference, http://www.thesedonaconference.org/

[30] The Sedona Conference® Best Practices Commentary on Search & Retrieval Methods (2007). http://www.thesedonaconference.org/publications_html

[31] The 2006 Socha-Gelbmann Electronic Discovery Survey Report, (2006). http://www.sochaconsulting.com/2006survey.htm/

[32] The 2008 Socha-Gelbmann Electronic Survey Report, (2008). http://www.sochaconsulting.com/2008survey.php/

[33] TREC: Text REtrieval Conference, http://trec.nist.gov/

[34] TREC Legal, http://trec-legal.umiacs.umd.edu/

[35] Weerkamp, W., Balog K., de Rijke, M. (2009) Using Contextual Information to Improve Search in Email Archives. In: 31st European Conference on Information Retrieval Conference (ECIR 2009), LNCS 5478, pages 400-411

[36] Zhao F.C., Oard, D.W., Baron, J.R. (2009) Improving search effectiveness in the legal e-discovery process using relevance feedback. In: Proceedings of the global E-Discovery/E-Disclosure workshop on electronically stored information in discovery at the 12th international conference on artificial intelligence and law (ICAIL09 DESI Workshop). DESI Press, Barcelona

Integreon™

Best Practices in Managed Document Review

February 2011

The key to minimizing the risks and maximizing the efficiency and effectiveness of the document review process is to construct, document and follow a defensible process based on best practices that reflect sound project management disciplines, good legal judgment, and counsel's specifications.

A Note on Legal Jurisdiction

This broad guide is intended to have specific application in discovery exercises in US jurisdictions but also to inform the practice surrounding disclosure exercises in UK and other former Commonwealth jurisdictions. Specific reference is made to recent revisions to the CPR regarding e-disclosure practice in the UK, in particular the new Practice Direction 31B (effective date October 1, 2010), which expands best-practice guidance for counsel engaged in litigation requiring electronic disclosure. Pending developments in electronic discovery/disclosure rules and procedures in Australia should be expected to align all three jurisdictions in respect to several elements, key being the requirement of maintaining defensibility.

In several respects, the UK Practice Direction incorporates principles of litigation readiness and e-disclosure best practice that have for far longer been the rule in the US under e-discovery amendments to the Federal Rules of Civil Procedure (FRCP), adopted in December 2006, and state law analogues. Without going into detail, under both the Federal Rules and PD 31B, parties and their counsel are obligated to confer regarding disclosure of electronic documents, and to agree on the scope of discovery, tools and techniques to be employed, and specifications for production (exchange) of documents, all with an eye to ensure cost-efficient and competent disclosure of relevant electronically stored information. And every process employed must be fully transparent and documented in order to contribute to a fully defensible discovery exercise in total.

As in the US, the adoption in the UK of a new definition of competent practice in the domain of e-disclosure can be daunting at first blush to many litigators, as it suggests a need for the lawyer to master a technological discipline somewhat alien to the traditional practice of law. Counsel is well-advised to seek competent e-discovery providers/partners to help navigate the e-disclosure landscape and to recommend processes and tools that have been proven in e-discovery practice.

Introduction

If there is a lot riding on the outcome of litigation, there is a lot riding on the manner in which discovery, and by extension, document review, is conducted.

Often we, both clients and counsel, think about conducting discovery and managing document review as necessary yet secondary concerns, a couple steps in priority and glory beneath the higher calling of designing and implementing case strategy. Moreover, in the past several years, there has been an increasing focus on cost-containment in this phase of discovery, leading to growing interest in simple and expedient solutions. But we should not lose sight of the stakes involved. Defensibility must remain the governing principle; we should want an efficient and effective process that meets a reasonable standard and that can be defended. In the wake of *Zubulake* and its progeny, specifically, we recognize that a defensible process, well-conceived and executed, is imperative *and* minimizes risk. Accordingly, counsel should guard against undervaluing discovery as a process. Best practice principles must be extended to the context of discovery and document review.

This paper outlines recommended best practices for managing document review – a basic best practice *guide* – having as its goals the design of an efficient, cost-effective and defensible workflow yielding consistent and correct work product.

Managed Document Review

There are, in basic construct, two standing industry models for outsourcing e-discovery document review projects – the managed review model and the staffing model.

Engaging a managed review provider is readily distinguishable from having a staffing agency supply temporary labor, such as attorneys or paralegals, to perform review. In the latter case, a law firm or client specifies reviewer qualifications and the staffing agency locates and vets the reviewers and assembles the team. But the staffing agency typically does little more than provide the raw workforce which must be trained, monitored, and supervised by counsel. In instances, the law firm or company also provides necessary infrastructure – physical space, technology systems, security controls, etc. – to support the review. Because staffing agencies don't assume responsibility for managing or governing the process, the law firm or client is solely responsible for planning, review design, assignment workflow, training, process documentation, reporting, and validation of results.

A managed review provider typically provides a review team, facilities, technical support, and project management, and shares with counsel responsibility for managing an efficient and defensible process. In the best examples, the review provider, whether a full-service e-discovery vendor or a stand-alone review operation, collaborates with counsel in recommending an optimal project workflow. In addition, the review provider offers proven operational features including complete metrics reporting to assist counsel in overseeing and ensuring an efficient and effective discovery exercise, from kick-off through post-production.

Whatever the choice counsel makes in selecting a review solution – whether review conducted by associates, by a temporary staff of contract agency attorneys, or outsourced to a managed review provider – the solution should reflect an approach steeped in an understanding of applied best practices.

The following sets forth a minimal, standardized, framework which can and should be adapted to meet the needs of specific cases.

PLANNING AND PROJECT MANAGEMENT

- Ensure a project plan is tailored to the specifications of counsel and consistent with best practices
- Deliver a key set of documents that govern the execution and project management of the review process

TEAM SELECTION AND TRAINING

- Develop specific job descriptions and define a detailed protocol for recruiting, testing, and selection
- Conduct reference and background checks, and a conflicts check, where necessary
- Employ team members previously used on similar projects
- Ensure the review team receives comprehensive substantive and platform training

WORKFLOW

- Design processes, assignments and quality assurance steps specifically geared to the project's requirements
- Demonstrate compliance with key security and quality standards while maintaining acceptable pace

QUALITY CONTROL

- Develop quality control processes to achieve key project goals
- Implement controls to manage privilege designation and preparation/ validation of results for production
- Test first review work product using sampling, targeted re-review, and validations searches
- Employ formal statistics to ensure the highest quality end result
- Maintain performance tracking for all reviewers

COMMUNICATION

- Develop a formal schedule of communications with counsel
- Calibrate initial review results, seeking counsel's guidance to confirm or correct results and to conform review protocol and training materials to insights gained

REPORTING

- Deliver regular, comprehensive reports to monitor progress and quality and to assist counsel in managing the review process

PRODUCTIONS AND PRIVILEGE LOGS

- Isolate and validate producible documents for counsel's imprimatur
- Prepare privilege logs in accordance with specifications set by counsel

POST-CASE

- Determine need for documents to be placed in a repository for future or related litigation
- Document the process from collection through production and assemble a comprehensive defensibility record

Best Practices

An effective document review team serves as a "force multiplier" that attempts, as closely as possible, to approximate the decisions that senior lawyers intimately familiar with the underlying case would themselves make if they had the time and opportunity to review each of the documents. A best-practices review establishes a construct in which counsel's guidance can be – and should be – assimilated into many discrete decisions across a team of reviewing attorneys and those many discrete decisions can be calibrated to deliver consistent results.

PLANNING AND PROJECT MANAGEMENT

There are two key objectives to the discovery process. The first is to identify documents – the production set – relevant to the matter at hand and responsive to the discovery request(s), with privileged documents held apart. The second is to recognize and bring to the attention of counsel the subset of documents that warrant particular attention, either because they support the case or are likely to be used by opposing counsel and therefore merit a prepared response. Achieving these goals requires sound planning and project management tailored to the directives from counsel.

The outcome of the planning process should be a set of documents that govern the execution and project management of the review process. These documents ensure that all the key elements of the project have been discussed and specify all decisions, tasks and approaches. The planning stage documents should include:

- Protocol plan
- Comprehensive project management manual
- Privilege review guidance notes
- Sample reports

The protocol plan documents the background and procedures for reviewing documents in connection with the specified litigation – it is a roadmap for the review team. A protocol plan typically includes a backgrounder to provide context for the review exercise (with information regarding the underlying litigation and a high level statement of the objectives of the review). Additionally, it includes detailed document review guidance, including a description of and examples of what constitutes relevance or responsiveness; how broadly or narrowly privilege is to be defined for purposes of review; what information or content is to be designated "confidential;" a primer on substantive issues that are required to be identified; and how other materials are to be treated, including documents that cannot be reviewed ("technical defects") and foreign language documents. The protocol also lays out the schematic or "decision tree" and procedures for how issues or questions are to be raised among the team members and with counsel.

The project management manual includes the review protocol and also lays out operational elements for the project, including: review scope; timeline; deliverables; staffing including team structure and job responsibilities, training, and work schedules; productivity plan; workflow; identification and features of the review application/platform; a quality control plan; feedback loops; query resolution process; communication plans, including reporting, validation methodology and final privilege review; and key project contacts, project closing procedures, and security protocols.

Privilege review guidance notes summarize guidance for the privilege review process and should cover the following areas: overview of reviewer roles, guidance on categories and the scope of privilege, guidance on accurate coding, and privilege logging of documents.

Sample reports provide counsel with examples of the reports that will be routinely delivered throughout the project. This is important to ensure up-front agreement on all reporting requirements.

TEAM SELECTION AND TRAINING

Assembling a review team entails formulating specific job descriptions, identifying the associated skill sets based on the parameters of the engagement, and defining a protocol for recruiting, testing, and selection.

The process should reflect relevant regulatory requirements and guidelines such as those set forth in ABA Formal Opinion 08-451, which states:

> *"At a minimum, a lawyer outsourcing services … should consider conducting reference checks and investigating the background of the lawyer or non-lawyer providing the services … The lawyer also might consider interviewing the principal lawyers, if any, involved in the project, among other things assessing their educational background."*

The level of training and experience of the review team is contingent upon the described task set. For example, a review for the purpose of redacting personal or confidential information may require limited legal training, and may be delegated to teams of paralegals under a lawyer's supervision. Other reviews require an exercise of judgment or discretion wisely entrusted to teams of qualified junior lawyers, or even elements of substantive legal knowledge within the purview of the most highly trained and experienced attorneys.

It is expected that all team members will receive thorough substantive training from counsel and an orientation or re-orientation to the selected review platform (application) prior to the commencement of each review. Early review results should be reported in detail to counsel and detailed feedback sought. In pro-actively soliciting counsel's guidance on any reviewed documents on which a question was raised by reviewers, and to confirm or correct coding decisions made early in the review, the

team is progressively more closely aligned to counsel's instructions. Review protocols should be fine-tuned or expanded, as necessary, as additional guidance is received from counsel.

WORKFLOW

Workflow design is a synthesis of science and art. How a reviewable population of documents is approached in review will determine both the efficiency and pace of review. Workflow on linear review platforms – using conceptual searching tools and clustering or similar technology – can be optimized by applying screens to a given document population, sorting into queues those documents having similar content or format from specified custodians, or isolating discussion threads. This can aid reviewers in making consistent calls more quickly. Additional techniques can be integrated into the process to speed review, including highlighting specific search terms within documents and segregating potentially privileged documents for focused review. Other techniques can be applied to ensure accuracy of review, such as employing a mix of sampling and validation searches and targeted re-review of reviewed documents, sampling of results by counsel, and employing a formalized query resolution process that requires counsel to formulate specific answers to questions in writing.

Workflow design includes the review tagging structure, incorporating desired behaviors, and constraints for individual tags. Consideration must also be given to the preferred treatment of document families, confidential or personal information, and whether redactions need to be applied.

Related issues include attention to data integrity and security protocols to be followed during review and on the review platform.

QUALITY CONTROL

Any endeavor involving human effort, employing tools designed by humans, is inherently prone to error. Therefore, the standard for discovery, or indeed execution of any legal service, is not perfection. Rather, work product is expected to be correct within tolerances defined at times as consistent with diligent effort and the exercise of reasonable care. The dividing line between inadvertent error and culpable error or wanton carelessness lies in whether *reasonable* care was exercised in avoiding and detecting such errors. For a specific example, Federal Rule of Evidence 502(b) provides that inadvertent disclosure of privileged material will not result in waiver where the holder of the privilege (through counsel) "took reasonable steps to prevent disclosure" and "promptly took reasonable steps to rectify the error." So one question is, how do we define *reasonable steps* [to prevent disclosure of privileged material] and, more broadly, *reasonable care*, in the context of document review?

Reasonable care, in this context, equates to what we call "defensible" – and requires, at a minimum, an intelligently designed suite of quality control measures matched to rigorous training, performance measurement, and reporting. A very capable quality control regime includes:

- Intelligent validation of results to ensure the set of reviewable data has been reviewed in its entirety by the appropriate reviewers
- Targeted review to detect potential errors and to identify materials requiring further review
- Targeted review to isolate from the production set all privileged documents

Quality controls should be implemented in at least two key areas: privilege designation and validation of presumptive production sets. Review results should be "tested" and determined to be:

- Consistent across the entire data set and team, across multiple phases of a project, and with protocol treatment for families, duplicates, etc.
- Correct in that it meets parameters for relevance, privilege, confidentiality, and issue coding, and that all potential privilege has been identified

There are significant challenges in designing and executing a rigorous and effective quality control regime. Where sampling is relied upon, there may be reason to employ statistical methods in order to identify statistically sound and representative random samples of a document population for re-review. The most effective and, arguably, more defensible approaches combine sampling with intelligently targeted quality control elements to identify documents meriting a second level of review, and also solicit continuous input from counsel to calibrate the review team. All quality control elements should be designed with counsel's input and documented.

COMMUNICATION

Best practices mandate developing a formal and regular schedule of reporting and communications among the review team, its managers, and supervising counsel throughout the process. During ramp-up, communications should be geared to ensure that supervising counsel is available to help confirm review guidelines and answer reviewer questions. A schedule of regular calls should be established to review progress and any issues. A best-practices communication plan will also document points of contact, escalation processes, and appropriate means of communication.

REPORTING

Reporting is a key element of the review process and is the primary means by which counsel is presented with information necessary to assess, in real time, whether a review is on track and on pace, how accurate the results are, the breakout of designations made for documents

reviewed thus far, and the number of interesting ("hot") or problematic documents. Review reports, issued at agreed-upon intervals, deliver invaluable information on productivity, accuracy, operational issues, technical issues, team structure, folders released, and other requested metrics. Good systems can now generate reports containing these and other data points automatically. Best practice requires, of course, that the review vendor and counsel actually *read* the reports and act on information gained.

PRODUCTION AND PRIVILEGE LOGS

Where production is to be made to an adversary or requesting agency, best practices necessitate counsel and vendor to agree well ahead of time on production specifications (principally, format and included fields) and procedures. The provider handling processing and hosting of reviewable documents should provide to counsel a comprehensive production log, cross-referencing production ID numbers (Bates numbers) to document ID numbers on the review platform and correlated to the original data collection. Privileged and redacted or withheld documents ordinarily would be logged by the review team or its managers, with the format and content of each log also having been agreed upon ahead of time. Final logs (and final privilege determinations) should be reviewed by counsel prior to production.

POST-CASE/DOCUMENTING THE PROCESS

Counsel should determine early in the process whether some or all documents should be maintained in a repository for future or related litigation, and necessary arrangements should be made with the responsible vendor. An advantage that can be gained through using a repository is that, once made, final privilege designations can be preserved if the same dataset is subject to future or related litigation discovery.

As a final element of best practices, counsel and vendors involved in all aspects of a discovery exercise, specifically including review, assemble a complete documentary record of the discovery process, including specifications of the collection, processing, review, and production(s). Such a record, which we refer to as a "defensibility binder," is a valuable tool for counsel as a historical record to answer questions raised at a later date and as a means of demonstrating that discovery was undertaken with diligence and reasonable care.

Conclusion

Document review is a critical, resource-intensive component of the e-discovery process that, in order to be successful, requires active and competent project management, following a suite of well-designed processes that reflect relevant and agreed upon best practices. The result is the timely and cost-effective delivery of defensible work-product that facilitates the overall litigation process and enhances the favorability of its outcome.

For more information contact:

Foster Gibbons
(foster.gibbons@integreon.com)

Eric Feistel
(eric.feistel@integreon.com)

www.integreon.com

Integreon™

ORRICK

Searches Without Borders

By Curtis Heckman, eDiscovery Associate,
Orrick, Herrington & Sutcliffe LLP

Even when working with native language documents, collection and production searches can be time-consuming and challenging. A multi-lingual environment introduces additional complexity to searching and therefore exposes the litigant to a greater risk of human and technology-based errors. This paper discusses the challenges of multi-lingual searches and provides simple recommendations for crafting defensible searches.

Computers were originally developed on an English-based system. Thus, the original assignment of alphabetic characters for computers was made in favor of Latin letters. Soon after, engineers developed computers using non-Latin alphabets. For a computer's purposes, a number is assigned to each alphabetic character. Strung together, these numbers are referred to as code pages. Communication between computers "speaking" languages based in different alphabets can become muddled beneath the surface as a result of their different code pages. Even different versions of the same software can vary in the interpretation and translation of non-Latin characters. Keyword searches that work in the language environment in which they were created may, despite looking identical on screen, miss documents created using a different alphabet. Accordingly, a comprehensive and accurate search may require that a litigant consider the original language environment of potentially-relevant documents.

Each custodian's software and hardware, and even the alphabet used to type a keyword, significantly impacts the accuracy of the search. If a litigant knows or anticipates that its universe of potentially-relevant documents contains information created in more than one alphabetic system, that party should consider using keyword variations in its search protocol.

The best way to ensure a defensible search is to have a quality assurance system in place to test the results of the search and collection methods at the outset.

- The Sedona Conference recommends parties evaluate the outcome of each search, using "key metrics, such as the number of included and excluded documents by keyword or filtering criteria, can be used to evaluate the outcome."[1]
- Ask questions during the initial custodian interviews to identify the language(s) used for both formal and informal communications. Also identify alternative alphabetic characters on the custodian's keyboard, and what, if any, type of software the custodian used to type in the computers non-primary language.
- Question the Information Technology department and any discovery vendor regarding the ability of proposed search engines to account for different character sets and code pages.

---

[1] Jason R. Baron et al., *The Sedona Conference: Commentary on Achieving Quality in the E-Discovery Process* (May 2009) at 15.

ORRICK

- If the production or review is to involve translations, determine whether the translation is to be a machine-based translation and whether the translator has the capability to differentiate mingling characters.
- Document the entire process underlying the determination and deployment of key words.

A receiving party should also be fully prepared to discuss the producing party's search obligations if multiple alphabetic systems are anticipated:

- Consider meeting with a consultant or expert who understands multi-language searches and productions prior to the initial meet and confer.
- Negotiate the parameters of the search at the meet and confer. Clearly identify and convey your production expectations.
- If no agreement is reached and the litigation moves to motions practice, have an expert provide technical affidavits regarding searching and production in multi-system environments.

Litigation imposes significant difficulties for international corporations, not the least of which is multi-lingual searches. Knowing the challenges before tackling cross-border searching and documenting each step establishes reasonableness and defensibility.

Submission by: Curtis Heckman of Orrick, Herrington & Sutcliffe, LLP.

Mr. Heckman is a member of Orrick's eDiscovery working group, which is Orrick's practice group devoted to eDiscovery. He is resident at Orrick's Global Operations Center ("GOC") in Wheeling, WV and can be reached at (304) 231-2645 or checkman@orrick.com. The GOC is Orrick's on-shore outsourcing center and home of Orrick's Document Review Services and Data Management Group.

Full biographies and description of the practice group are available at http://www.orrick.com/practices/ediscovery.

ORRICK

Position: The discovery process should account for iterative search strategy.

By Logan Herlinger and Jennifer Fiorentino,
eDiscovery Associates, Orrick, Herrington & Sutcliffe LLP

Developing an informed search strategy in litigation, whether it involves key terms, technology, data sources, document types or date ranges, is an iterative process. Unfortunately, many judges and attorneys still approach discovery with outdated perceptions of search methodology. Litigants should not be permitted to submit a keyword list of fifty terms for the opposing party to use in their document review and production, and expect that it will define the scope of discovery. Nor should they be allowed to dictate which data sources the other side should search. Members of the legal community need more education on recent advancements in search strategy to appropriately define the scope of relevant information. An informed search strategy leads to more efficient and accurate responses in discovery, but the parties must meet and confer often and in a timely fashion to take advantage of these advancements.

For years, members of the legal profession have called attention to the problems inherent in the use of keyword search terms. Attorneys may draft search terms too narrowly or too broadly. Some terms may yield a large amount of hits unexpectedly, such as a term that is automatically generated in every company email signature. These problems exist because there is significant variance in the way individuals use language, and it is impossible to fully predict that use when first drafting search terms. Similarly, attorneys must focus their keyword searches on the appropriate data sources. Otherwise, they waste time and money on irrelevant information.

To address these issues, parties need to test the search terms' performance repeatedly in an iterative process to determine how they interact with the data set at issue. As data sets increase in size, search term problems are exacerbated, and more time must be spent testing the terms to determine their effectiveness. Cooperation between the parties is essential to make this process as efficient and effective as possible. Attorneys must take the time to become familiar with their client's data, conducting custodian interviews to learn about unique use of language and to determine where relevant data is most likely to reside. Each side should bring this knowledge to the meet and confer. The parties should discuss and agree upon an initial search strategy. Thereafter, the parties should test the chosen methodology and meet often and in a timely fashion to discuss the results. Because of the disparity in understanding of search methodology amongst judges and attorneys, it is often appropriate to include a third party data analytics provider to assist in the process and, if needed, defend its use before the court.

In order to effectively implement an iterative search strategy, the courts and parties must account for the time involved in testing, meeting, and applying the search terms in the document review process. It is widely recognized that standard linear document review is the number one

driver of discovery cost and time. While an iterative search strategy takes time upfront, its use should yield a focused and accurate data set for attorney review, which significantly limits the review costs and increases efficiency in responding to discovery requests.

There is prior literature that begins to address this position. In 2008, The Sedona Conference published the "Cooperation Proclamation," which addresses the increasing burden that pre-trial discovery causes to the American judicial system and champions a paradigm shift in the way the legal community approaches discovery. *See* The Sedona Conference, *The Sedona Conference Cooperation Proclamation* (2008), http://www.thesedonaconference.org. Although still zealously advocating for their clients, attorneys should cooperate as much as possible with the opposing party on discovery issues, particularly those that relate to ESI. This reduces costs for the clients and allows the attorneys to focus on the substantive legal matter(s) at issue, which in turn advances the best interests of the client. Along the same lines, there has recently been a push for proportionality and phased approaches to eDiscovery. The Sedona Conference Commentary on Proportionality in Electronic Discovery says that it may be "appropriate to conduct discovery in phases, starting with discovery of clearly relevant information located in the most accessible and least expensive sources." *See* The Sedona Conference, *The Sedona Conference Commentary on Proportionality in Electronic Discovery* at 297 (2010), http://www.thesedonaconference.org. This work supports the value of an iterative search strategy and for parties to meet and confer often and timely on the issue.

Attorneys must be prepared to educate the judiciary on the importance of a fully developed search strategy. They have to explain that while the iterative search process and subsequent meet and confers take time, the cost savings and increased efficiency in pre-trial discovery are significant and well worth the effort.

Submission by: Jennifer Fiorentino and Logan Herlinger of Orrick, Herrington & Sutcliffe LLP.

Ms. Fiorentino and Mr. Herlinger are members of Orrick's eDiscovery Working Group. Ms. Fiorentino is resident in Orrick's Washington, D.C. office, and can be reached at (202) 339-8608 or jfiorentino@orrick.com; Mr. Herlinger is resident at Orrick's Global Operations Center ("GOC") in Wheeling, W.V. and can be reached at (304) 234-3447 or lherlinger@orrick.com. The GOC is Orrick's on-shore outsourcing center and home of Orrick's Document Review Services and Data Management Group.

Full biographies and a description of the practice group are available at http://www.orrick.com/practices/ediscovery.

Adaptable Search Standards for Optimal Search Solutions

Amanda Jones
Xerox Litigation Services
Amanda.Jones@xls.xerox.com

ICAIL 2011 Workshop on Setting Standards for Searching Electronically Stored Information in Discovery Proceedings – DESI IV Workshop – June 6, 2011, University of Pittsburgh, PA, USA

The goal of this year's DESI IV workshop is to explore setting standards for search in e-discovery. Xerox Litigation Services (XLS) strongly supports the effort to establish a clear consensus regarding essential attributes for any "quality process" in search or automated document classification. We believe in the principles of iterative development, statistical sampling and performance measurement, and the utilization of interdisciplinary teams to craft sound information retrieval strategies. These will strengthen virtually any search process. Still, XLS also recognizes that there is no single approach to search in e-discovery that will optimally address the needs and challenges of every case. Consequently, there cannot be a single set of quantitative performance measurements or prescribed search protocols that can reasonably be applied in every case. Instead, we agree with the authors of "Evaluation of Information Retrieval" (Oard et al. 2011) that the discussion of standards for search should concentrate on articulating adaptable principles, clear and concrete enough to guide e-discovery practitioners in designing search solutions that are well-motivated, thoroughly documented and appropriately quality-controlled, with the flexibility to allow creative workflows tailored to the goals and circumstances of each matter.

Because of the unique and complex challenges ever-present in search in e-discovery, XLS would contend that the key to designing successful search strategies is the ability to explore multiple perspectives and experiment with a variety of tactics. Countless factors influence the quality of automated search outcomes. Therefore, it will be vital to the advancement of search techniques to adopt standards that encourage research on the sources of variability in search performance and create the latitude that is needed for ongoing hypothesis-testing and midstream course correction.

One source of variability in text-based search performance that XLS has already identified and addressed is data type. Relevance is manifested in markedly different linguistic patterns across various types of documents. So, XLS has elected to utilize distinct classification models for spreadsheet data, email data, and other text-based data for most projects. Developing and implementing distinct models for these three classes of data requires an additional investment of time and resources, but has consistently translated into significant performance gains for the population as a whole. So, it is the approach that we currently use to mitigate this source of performance variation and ensure the highest possible quality in our automated search results. Our research into this is continuing, though, and we are open to adopting a new equally effective and less labor-intensive tactic for managing linguistic variation across-data types.

Both within and outside Xerox, research in machine learning, information retrieval, and statistical data-mining is progressing rapidly. Thus, it is important to not only to devise creative solutions to known sources of variation in search performance, but also to have the freedom to explore the full potential of emerging automated search technologies. XLS is currently experimenting with ways to optimize search results by utilizing multiple techniques and technologies simultaneously, incorporating input from all sources that enhance the final results. In our observations, combining search tactics often leads to significantly higher performance metrics than can be achieved by any of the individual tactics alone. In one preliminary investigation across several matters, for example, we found that combining scores from one statistical algorithm applied to the metadata of a population with scores from a completely different statistical algorithm applied to the full text of the population consistently increased both precision and recall.

Similarly, we have also found it constructive to treat certain responsive topics or data types within a project with one search technique while using alternative approaches for other topics or data sources. For example, by analyzing patterns of error generated by our statistical algorithms, it has been possible for us to identify opportunities to use highly targeted linguistic models to correct those errors in the final result set. In general, our experimentation with hybridized search strategies has proven extremely fruitful and there are many avenues of investigation left to pursue in this area. This is a major motivating factor behind XLS's support of standards that would promote the novel application of any combination of available search resources, provided the efficacy of these applications were adequately demonstrated.

Obtaining a better understanding of the limitations of various search techniques is just as important as exploring the potential of new search technologies, because the limitations will also engender adaptive search strategies. Any text-based automated classification system will be subject to certain dependencies and limitations. For example, achieving comprehensive coverage with a high degree of accuracy is often challenging for search systems that rely on linguistic patterns to identify responsive material when responsive documents are "rare events" in the data population – primarily because there are simply fewer examples of the language of interest available to generalize. So, each and every responsive document is more noticeably impactful in the final results and performance metrics. In a case like this, more data is generally needed to achieve high precision and recall. It is sometimes possible, though, to mitigate the need for additional data utilizing linguistic and/or statistical approaches to increase the density of responsive material in a subset of the data population thereby increasing access to responsive linguistic material for generalization. Even then, though, it may require significant extra effort and ingenuity to ensure accurate and comprehensive coverage of the topic.

Further, the rate of responsiveness in a population interacts in a complex way with the definition of the responsive topic itself to influence the level of difficulty that can be anticipated in the development of a successful search strategy and the extent to which special tactics will need to be pursued. While it is not often discussed in great detail, it is extremely important to consider the subject matter target for a case when assessing options for search strategy. The way in which responsiveness is articulated in requests for production can have a profound impact on search efficacy. For example, all of the following subject

matter attributes will play a role in shaping the inherent level of difficulty in using automated search techniques to evaluate a population for a given topic:

- Degree of subjectivity – e.g., a request for production may specify that all "*high level* marketing strategy" documents should be produced, but an automated search approach will likely struggle to differentiate between documents that constitute "high level" discussions and those that represent "routine" marketing conversations
- Conditions on modality – e.g., a request for production may specify that all "*non-public* discussions of pricing" should be produced, but linguistic distinctions between private and public conversations often prove unreliable causing automated approaches to confuse pricing discussions between corporate employees with similar discussions appearing in the media, etc.
- Linguistic variability – e.g., a request for production may specify that all "consumer *product feedback*" should be produced, but consumer feedback may touch upon any number of product features, may be positive or negative, may appear in formal reports or informal emails, and may be expressed in any number of unpredictable ways that could prove challenging for automated search systems to capture comprehensively
- Linguistic generalizability – e.g., a request for production may specify that all "negotiations with *retailers*" should be produced, but if the corporate entity routinely deals with thousands of retailers, it would be difficult, if not impossible, for an automated search system to successfully recognize the complete set of potentially relevant retailers and differentiate them from entities such as wholesalers or suppliers, etc.
- Conceptual coherence – e.g., a request for production may specify that all "discussions of *product testing*" should be produced, but if this is intended to include R&D testing, Quality Control testing and Market Research testing, then there will actually be three distinct concepts to capture, each with its own community of expert speakers with unique jargon and communication patterns such that capturing all of these sub-topics equally successfully may challenge automated search systems

These factors interact not only with rate of responsiveness but also with one another to shape the target of the search effort. Analyzing the subject matter of a case to identify attributes that may introduce difficulties for automated search will make it possible to devise methods for overcoming the challenges.

There are, in fact, numerous options for coping with the various situations highlighted above. Sometimes the solution will be as simple as choosing one search technique over another. At other times, it may be most effective to collaborate with the attorney team to operationalize the definition of responsiveness to minimize the need for subjective interpretation or fine-grained subject matter distinctions. At other times, the best choice may be to create distinct models for the most critical sub-topics in an especially wide-ranging request for production to ensure that they will receive ample effort and attention, reducing the risk of having their performance obscured by the search results for other more prevalent topics. Undertaking a preliminary subject matter analysis and consultation with the case team, along with early sampling and testing in the corpus, will typically enable the formulation of a

project proposal that will provide value for the client while accommodating the realities of the search situation.

Finally, while much of the above discussion has centered on the use of in-depth analysis and a multitude of search tactics to achieve the highest possible quality results, XLS acknowledges this level of analysis and investment of expert resources is not always feasible. In fact, it may simply be unreasonable given the practical constraints of the case or its proportional value to the primary stakeholders. Open and frequent communication with the attorney team and client for the matter will not only enhance the quality of the subject matter input for the project, but also afford them opportunities to contribute their invaluable expert opinions regarding the reasonableness of the search for the matter at hand.

In sum, XLS adopts the position that search results in e-discovery should be judged relative to the goals that were established for the project and that the search process, rather than the technology alone, should be scrutinized. We recognize it would be advantageous to have a single concretely defined protocol and technology applicable to every matter to achieve high-quality results quickly, cheaply, and defensibly. However, it would be naïve to suggest the unique topics, timelines, resources, parties, data sources and budgetary constraints associated with each matter could all be treated successfully using the same search strategy or the same quantitative measures, especially when current technologies are in a state of growth and evolution. It does a disservice to both the complexity of the problem and to the value of human insight and innovation in tailoring custom solutions to adapt to specific needs.

discovery
ie discovery, inc.

ISO 9001: A Foundation for E-Discovery

As the e-discovery industry strives for common standards and practices, an ideal solution exists: ISO 9001.

by Chris Knox, Chief Information Officer, IE Discovery
and Scott Dawson, President of Core Business Solutions

Executive Summary

A lack of standards and common practices hampers the processing of Electronically Stored Information (ESI) for litigation. Industry thought leaders, as evidenced by recent work by the Sedona Conference, Text Retrieval Conference (TREC) Legal Track, as well as this DESI meeting, are actively seeking standards to define and manage the process of discovery. This is necessary for both competitive differentiation in the marketplace, as well as to satisfy a growing demand for transparency and documentation of the Discovery process from the judiciary.

The legal profession has focused much of its energies seeking benchmarks and standards in the search process, but there is a need to be able to certify repeatable, defensible, and consistent business processes through the entire e-discovery process. In many of the ongoing industry discussions, the ISO 9000 standards family arises as one of the ideal models for how to provide certification and standardization. In fact, we believe that the ISO 9000 family of standards is not just a model, but is ready today to provide a common standard of quality for e-discovery. In addition, the model provides a framework for an industry-specific solution that can emerge to solve the growing complexity and difficulties found in e-discovery.

Introduction

The Discovery Management industry does not have a defined baseline for quality. Complicating matters, the discovery of evidence for litigation was, until relatively recently, a paper-based process. As such, any existing industry standards and quality expectations are still primarily paper-based or focused on standards of quality control for scanning paper documents to digital formats. As the industry has adapted to process the exploding universe of digital media, and new products and processes are introduced, no new set of quality standards has emerged specifically to govern the discovery of Electronically Stored Information (ESI).

In other industries, standards help inform buying decisions, provide a common language and point of reference to communicate quality. When purchasing in a manufacturing vertical (such as automotive and pharmaceuticals), buyers can expect a baseline of quality based on certifications. The e-discovery services industry is largely cost driven, with buyers purchasing e-discovery services as if it were a commodity but without the means to ascertain the level of quality they can expect. However, the purchasers of e-discovery legal services cannot expect quality service at every price point because of the lack of accepted industry practices.

Industry standards are not simply a marketing tool to sell services; standardization of processes is an explicit requirement from the judiciary.[1] Primarily in the area of search technology, courts have confirmed that standards are necessary for establishing defensible e-discovery practices. In addition, the Federal Rule of Civil Procedure 26(g)(1) requires attorneys to certify "to the best of the person's knowledge, information, and belief formed after a reasonable inquiry" that disclosures are "complete and correct."

[1] *William A. Gross Construction Associates, Inc. v. American Manufacturers Mutual Insurance Co.*, 256 F.R.D. 134, 134 (S.D.N.Y. 2009) ("This Opinion should serve as a wake-up call to the Bar in this District about the need for careful thought, quality control, testing, and cooperation with opposing counsel in designing search terms or "keywords" to be used to produce emails or other electronically stored information")

We believe these requirements can be satisfied with the adoption of quality management and documentation processes in the e-discovery industry.

The discussion at this conference[2] and others like it underscores the growing consensus that quality standards are necessary in creating the common baseline for defensible and standardized e-discovery practices. However, the industry is just beginning to grapple with issues such as the criteria used to search electronic records for responsive documents. For example, the Text Retrieval Conference (TREC) Legal Track is evaluating the effectiveness of various methods of information retrieval technology to find a baseline quality expectation[3] Similarly, the Sedona Commentary on Achieving Quality in E-Discovery calls for development of standards and best practices in processing electronic evidence.[4] These efforts are feeding a larger effort to create defensible standard practices for the industry.

Most professionals in the e-discovery industry understand that there will never be a comprehensive e-discovery process; the demands of searching, reviewing, and producing evidence from the complex, diverse, and ever-expanding universe of discoverable data ensures that standardization will likely be impossible. An even bigger obstacle is the rapid changes in technology. For example, the use of advanced information retrieval technology to augment the human review process is constantly evolving.[5] Also, protocols are case-based; what may be a perfect solution in one situation may not be appropriate for the next.

ISO 9001 is an ideal solution for this state of affairs because it is designed to deliver the best solution for different situations. Because ISO 9001 is a baseline standard, it is flexible enough to address this complex challenge as few other approaches can. ISO 9001 has been held up as a standard that is a useful example of the type of standard the e-discovery industry can hope to develop. We believe that ISO 9001 is in fact not just an example, but a workable, real-world solution that provides a solid foundation for the e-discovery industry today.

What is ISO 9001?

The ISO 9000 family of standards is an internationally accepted consensus on good quality management practices. ISO 9001 is an international quality certification that defines minimum requirements for a company's Quality Management System (QMS). A company's QMS includes the organization's policies, procedures and other internal requirements that ensure customer requests are met with consistency and result in customer satisfaction. Some of the areas of an organization within the scope of ISO 9001 include:

- **Customer contracts**
- **Hiring and employee training**
- **Design and development of products and services**
- **Production and delivery of products and services**
- **Selection and managing of suppliers**

[2] In Search of Quality: Is It Time for E-Discovery Search Process Quality Standards? Baron, Jason E-Discovery Team blog. (http://e-discoveryteam.com/2011/03/13/in-search-of-quality-is-it-time-for-e-discovery-search-process-quality-standards/)

[3] J. Krause, Human-Computer Assisted Search in EDD, Law Technology News, December 20 (2010).
and Oard, et. al. Evaluation of information retrieval for E-discovery, Artificial Intelligence and Law, December 22 (2010).

[4] The Sedona Commentary on Achieving Quality in E-Discovery, May 2009. Principle 3. Implementing a well thought out e-discovery "process" should seek to enhance the overall quality of the production in the form of: (a) reducing the time from request to response; (b) reducing cost; and (c) improving the accuracy and completeness of responses to requests.
The type of quality process that this Commentary endorses is one aimed at adding value while lowering cost and effort.

[5] Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11 (2011).

To maintain the certification, an organization must implement:

- **Management responsibilities**
- **Internal quality audits**
- **Monitoring and measuring**
- **Continual improvement**
- **Corrective and preventive actions**

To receive an ISO 9001 certification a company must put the required QMS processes and controls in place, monitor performance of its processes and demonstrate continual improvement. Many companies hire an experienced consulting firm to assist with these preparations. Once the QMS is in place, a registrar (or certification body) is hired to audit the company's compliance with ISO 9001 requirements. If discrepancies are found during the audit, they must be corrected before the ISO 9001 certificate is issued. One of the most demanding aspects of the ISO 9001 certification is that it must be maintained through regular audits (bi-annual or annual) conducted by the selected registrar.

To maintain certification, organizations must provide measurable targets for improvement and data to show current and past performance. This information is kept in a quality manual, a general description of how a company operates and how meets ISO 9001 requirements. An organization provides specific procedures or work instructions determined by the management as needed to ensure processes meet the stated quality objectives.

In addition, an organization must maintain historical records that demonstrate compliance with company procedures and the ISO 9001 standard and train employees and management in the required responsibilities, ISO awareness and understanding of the quality policy, administrative procedures, and the audit process. Customer feedback is another essential component, which demands tracking customer complaints, compliments, and overall satisfaction. A management representative is assigned to coordinate the ISO program and a regular management review meeting should assess the progress and initiate improvements as needed. In addition, a team of employees trained to conduct an audit similar to the registrar's audit must conduct a formal internal audit, on top to the annual outside auditor review.

Through these requirements, organizations will likely find that they have to ensure that rigorous documentation of processes is in place. And because the program demands continual review and process improvement, the certification makes certain that an organization's services, documentation, and processes are consistently updated and streamlined.

The Benefits of ISO

From an organizational standpoint, adopting and adhering to an ISO 9001 compliant QMS creates a more organized operating environment, attracts new customers, and generally leads to a higher level of satisfaction among those customers. For e-discovery practices, the certification process would demand documentation and policies be put in place that are available as a reference or even supporting materials that attest to an e-discovery vendors good faith efforts to provide the highest standard of care in litigation.

From a practical standpoint, the certification forces an organization to continually upgrade and reconsider its processes. In outlining any current operations, organizations must add the requirements of the ISO 9001 standard and optimize processes, meaning internal operations can be quickly enhanced and streamlined. And, as noted, after achieving certification, the process mandates continual process improvements. A recent survey of 100 registered firms reported the average improvement in operating margin at 5 percent of sales. These firms also reported faster turnaround times, and a reduction in scrap and overtime.

In addition, the ISO process facilitates increased quality awareness. During implementation, quality awareness will increase, since all staff must be trained on ISO 9001. The QMS will also demand built-in systems to report on key quality indicators, which will significantly reduce the reoccurrence of problems. This helps develop a strong quality culture, where the staff recognizes problems such as systems or process issues and work on fixing them, rather than placing blame with an individual. And with ISO 9001

certification, employees learn processes more quickly and reduce misunderstandings with customers. If a problem does occur, it is traced to its root cause and fixed.

While there is no accepted ISO certification requirement in the e-discovery industry, ISO 9001 certification is becoming a requirement to do business in many markets. We believe that as sophisticated business enterprises bring the e-discovery process in-house and away from law firms, the expectation of ISO certifications will increase. A recent survey of ISO 9001 certified companies shows that 41 percent were asked to achieve certification by a client. Considering that it can take 6 months or longer for some organizations to achieve certification, already having a compliant QMS in place can be a distinct advantage. E-discovery vendors that do adopt the standard now, ahead of any possible requirement to do so, have a distinct marketing advantage, as they are able to declare their processes conform to an internationally recognized standard that few competitors can claim.

The ISO Organization

Perhaps the most important benefit of ISO certification is the broad, international acceptance the standard has achieved. The International Standards Organization is a combination of the national standards institutes in roughly 157 countries. The ISO 9000 family of international quality management system standards is perhaps the best known example of the organization's output, but it is only one of the many standards produced.

The ISO 9000 standards provide a basis for certifying compliance by individual organizations with standards in the family. An e-discovery company may qualify for basic ISO 9001 certification, or, more optimally, an industry-specific standard could be created to provide applicable certification to their operations in this field. And when a company or organization is independently audited and certified to be in conformance with any ISO 9001 standard in the family,that organization may also claim to be "ISO 9001 certified."

What ISO Does and Does Not Do

The ISO 9001 certification is distinct because it demands patience and an ongoing process of improvements. Other standards offer a regimen of self-help and implement more advanced management techniques in an organization. But as management and staff turnover naturally occurs, organizations lose interest and forget what they are working on without the ongoing commitment to the ISO 9001 audit process.

ISO mandates that an organization's management has a defined quality standard and meets these goals. Compared to the ISO model, other certification processes are often static. Once certified an organization can claim to have achieved the standard, but there is no required maintenance. For example, the Capability Maturity Model Integration (CMMI) in software engineering makes similar demands, but without demanding process improvement or providing a point of reference for appraising current processes.

Of course, no certification guarantees quality service; rather, they can only certify to potential customers that formal processes for measuring and controlling quality are being applied.

The ISO 9001 Family

Figure 1: ISO 9001 and its industry variants

ISO 9001:2008

Industry-specific standards:

AS9001	Aerospace Industry Standard
ISO/TS 16949	Automotive Industry Standard
TL 9000	Telecom Industry Standard
ISO 13485	Medical Industry Standard
ISO/TS 29001	Petroleum, petrochemical and natural gas industries Standard.
ISO 17025	Calibration and Test Laboratories
ISO 22000	Food Safety

Related management-system standards:

ISO/IEC 27001	Information security management
ISO/IEC 20000	IT service management
ISO 14001	Environmental management standards
ISO 26000	Social responsibility
OHSAS 18001	Occupational Health and Safety

As the chart above indicates, a number of industries have created ISO 9000 variants with specific requirements. Most are in manufacturing fields, although the model can certainly be adapted to create a standard specific to the processing of ESI. In addition, management system standards have created systems for implementing international standards for social responsibility, as in the ISO 260000 model or the environmental safety model defined by ISO 14001.

Of particular interest to e-discovery service providers, the ISO 27000 standard is designed to identify and manage risks posed to business information by data theft or accidental loss. It provides guidelines for putting a secure infrastructure in place and implementing a risk management process and corporate policy to minimize data loss. This is the one existing ISO standard e-discovery vendors can and should actively consider adopting in addition to the ISO 9001. E-discovery service providers can have their processing centers certified under ISO/IEC 27001 certification as an assurance to customers that any ESI handling and processing is done with a commitment to security and data integrity.

Litigation and support suppliers will certainly benefit from the adoption of the general ISO 9001 standard. However, many industries have benefited from the adoption of an industry-defined subset of 9001. These subsets were all proposed and developed by professional organization and industry experts with the intent of addressing perceived weakness in ISO 9001 relative to that specific industry. Because a number of initiatives and projects are underway that attempt to define and create a framework for acceptable e-discovery practices, these efforts could certainly be used to jump-start an effort to define an e-discovery ISO 9001 model.

ISO 9001 and the E-discovery Industry

Industry organizations have begun to make some initial attempts at creating standards and best practices. The Sedona Conference has a number of guides and best practices recommendations available for e-discovery topics, including search protocol and choosing an e-discovery vendor.[6] The Electronic Discovery Reference Model (EDRM) has led an effort to create a standard, generally accepted XML model to allow vendors and systems to more easily share electronically stored information (ESI).

[6] The Sedona Conference Publications (http://www.thesedonaconference.org/publications_html)

However, industry best practices are currently only recommendations, and technical standards such as the proposed XML schema are most useful in creating consistent standards and attributes for products. ISO focuses on how processes work and how work product is produced and is not a technical or product related standard. Technical or product related standard certifications are generally hard to come by and are most useful only for technology vendors and not service providers.

As noted, the ISO 9001 standard is a general, baseline and provides only high-level guidance. A number of industry sectors have created standardized interpretations of the ISO guidelines for processes in industries as diverse as aerospace manufacturing and medical devices. Industry-specific versions of ISO 9000 allow for industry-specific requirements, and allow for the training and development of a base of industry auditors that are properly qualified to assess these industries. Standardizing processes could standardize pricing as well – or at least create a common language for pricing e-discovery services.

Courts have repeatedly found that a failure to adequately document the steps taken to sample, test, inspect, reconcile, or verify e-discovery processes is unacceptable and can result in court-imposed sanctions.[7] The profession may resist applying metrics to litigation as are applied in manufacturing and other industries, but the discovery phase of litigation is a business process, and a quantifiable one. There will always be questions of law in the discovery process that require a lawyer's judgment and discretion, but within the process, service providers can and should apply some of the same rigor and standardization of service as seen in other industries. For example, some of the possible quality metrics that can be measured are:

- **Defects per reviewed document delivered**
- **Search expectations- how many images, graphics, or embedded documents were successfully indexed**
- **The error rate for files loaded to a repository of the total number of files received**
- **Deadlines met or missed**
- **A measure of data collected which was ultimately deemed non-relevant**
- **Search accuracy and recall**
- **The number of corrupted files loaded to a repository prior to review**

In order to implement such standards, definitions must be agreed upon. For example, such foundational issues such as what is a document and what is a container file. The ongoing research by TREC and other technical studies can continue to develop baseline measures for successful e-discovery search and document review. These measures and metrics should then be considered within the ISO 9001 framework to provide a baseline for quality of services.

Moving Forward

Organizations such as the Sedona Conference and the EDRM are two obvious candidates for promoting further efforts in this area. The ISO 9001 standard would in fact be an ideal vehicle for implementing the work these and other organizations done into search methodology and information handling across the industry. And together with the more detailed efforts to define and create best practices for the industry, perhaps an ISO 9001 standard for the management and handling of ESI can be formulated.

The primary driver for an e-discovery-specific ISO standard will be to ensure that when a customer purchases services from a certified source, they can have a level of assurance that the vendor has basic quality control practices in place. Most importantly, the ISO 9001 certification standard provides a third-party independent auditor who reviews the company's standard against the certification. Buyers do not want to trust a vendor with their data sets only find out a vendor does not have basic quality control measures in place.

ISO 9001 is a standard that may become necessary just to compete. The e-discovery industry can only stay fragmented for so long. In order to mature, the e-discovery industry needs a common language to both satisfy the demands of its customers as well as the growing chorus of judges and legal scholars looking for measurable quality standards.

[7] *The Pension Committee of the University of Montreal Pension Plan, et al. v. Banc of America Securities LLC*, et al., No. 05 Civ. 9016 (S.D.N.Y. Jan. 15, 2010)

About the Authors

Chris Knox, ***Chief Information Officer of IE Discovery***
Chris Knox has more than 16 years of project and resource management experience. He is responsible for the implementation of strategic initiatives and IT expenditures, as well as the development of company-wide operating procedures. Chris graduated from the University of Texas at Austin with a degree in Engineering and received his MBA from Syracuse University.

Prior to IE Discovery, Chris designed and implemented data collection networks for software developers and Fortune 500 corporations. Chris also previously created Geographic Information Systems for large municipalities, specializing in the development of algorithms for hydraulic models.

Scott Dawson, ***President of Core Business Solutions***
Scott has more than 20 years' experience in manufacturing with the past 10 years consulting with organisations seeking IS0 9001 certification. Scott is also an active voting member of the US Technical Advisory Group (TAG) to ISO Technical Committee 176 (TC 176), which is responsible for drafting ISO 9001 and ISO 9004 on quality management systems.

A Call for Processing and Search Standards in E-Discovery

Sean M. McNee, Steve Antoch
FTI Consulting, Inc.
925 Fourth Ave, Suite 1700
Seattle, WA 98104 USA
{sean.mcnee, steve.antoch}@fticonsulting.com

Eddie O`Brien
FTI Consulting, Inc.
50 Bridge Street, Level 34
Sydney, Australia 3000
eddie.obrien@ftiringtail.com

ABSTRACT
We discuss the need for standardization regarding document processing and keyword searching for e-discovery. We propose three areas to consider for standards: search query syntax, document encoding, and finally document metadata and context extraction. We would look to encourage search engine vendors to adopt these standards as an optional setup for the application of e-discovery keyword searches. We would encourage search engine users to apply these standards for e-discovery keyword searching.

Keywords
E-Discovery, search, engine, keyword, standards.

1. INTRODUCTION
E-Discovery document analysis and review continues to consume the bulk of the cost and time during litigation. As the e-discovery market matures, clients will have increased expectations about the quality and consistency of how their documents are collected, processed, and analyzed. It is also our assumption that e-discovery vendors will compete based on the quality and breadth of their review and analytic services offerings.

Seeing this as the changing landscape of e-discovery, we propose in this paper that the vendors of e-discovery software and services are encouraged to create and apply a set of shared e-discovery standards for document processing and keyword search. We hope that these standards would be organized and maintained by a standards committee such as the Sedona Conference [1] or follow the example of the EDRM XML standard.

2. AREAS FOR STANDARDS
We think there are several areas where consistency, speed, and quality could be improved by having an open and agreed to set of standards.

2.1 Search Query Syntax
Different information retrieval/search engine systems use different and often incompatible syntax to express complex searches. This can cause confusion for attorneys, for example, when they are negotiating search terms during Meet and Confer, or when they are trying to express a complex query to an e-discovery vendor.

A position paper at ICAIL 2011 Workshop on Setting Standards for Searching Electronically Stored Information on Discovery (DESI IV Workshop)

Examples of some difficulties worth noting:

- **Wildcard operators**. Should such operators match on 0 characters or not? For example, would (Super*FunBall) hit on both the SuperFunBall and SuperHappyFunBall, or only the latter?
- **Stemming and Fuzzy Searching.** Different IR systems provide support for different algorithms for term stemming and fuzzy searching (e.g. Porter stemming or Levenshtein distance). Attempting to standardize them might be too difficult in a standard. This would be an example of a value-add that a particular vendor could offer, but only of the lawyer understand and approve it.
- **Morphology and Word-breaking.** Concepts and word breaks are hard to determine in some languages. For example, Arabic has many ways to express a single term; Chinese and Japanese have ambiguous word boundaries.

These are only a few examples of the potential problems encountered when standardizing query syntax.

Our goal here is not to suggest that any given syntax is better than another. Nor is it to "dumb down" syntax by removing extremely complex operators. Rather, we see it as a chance to set a high bar as to what lawyers can expect from search engine systems in an e-discovery context. It is quite possible that some systems simply will not have enough functionality to support a standardized syntax. In this case, the lawyers are better off knowing of this limitation before e-discovery begins!

While the syntax varies by vendor, many complex expressions have direct correlations—there should be a mapping between them. Ideally mappings would make it possible to start with a standard syntax and have each vendor map the query to their equivalent native syntax. The standard syntax should be vendor-neutral; perhaps XML or some other formal expression language should be used to define it.

2.2 Encodings and Special Characters
Textual characters are encoded in documents through the use of various character sets. The first and most well-known character set is the ASCII character set describing 127 characters (letters, numbers, and punctuation) used in English.

Lawsuits, however, are language agnostic. Unicode [2] is the preferred standard from the ISO to represent a universal character set. To state that Unicode should be used as the standard encoding for all documents in e-discovery seems obvious—so, we should do it. What is not as obvious is the need for standardized set of test documents to validate the conversion to Unicode from a variety of data formats common to e-discovery.

Finally, the standardized search query syntax discussed above needs to be able to express searches for all Unicode characters, including symbols such as the Unicode symbol for skull-and-crossbones (0x2620):  .

2.3 Metadata and Content Extraction

A very small minority of documents in litigation are raw text documents. Most are semi-structured documents, such as emails, Microsoft Office documents, Adobe PDF documents, etc. These documents contain raw textual data, metadata, and embedded objects, including charts, images, audio/video, and potentially other semi-structured documents (e.g. a Microsoft Excel spreadsheet embedded in a Microsoft Word document).

We have an opportunity now to extend what has already been done in the EDRM XML standard to define what metadata should be considered standard extractable metadata for various file types. If we know in advance what is required, then we can ensure higher quality. For example, it will be easier to detect corrupt files. By standardizing, we also make meet-and-confer meetings smoother, as metadata no longer becomes a point of contention—both sides assume the standard is available.

2.3.1 Known Document Types

For known document types, such as Microsoft Office documents, there are several generally accepted ways of extracting content and metadata. These generally rely on proprietary technology, some of which are free (Microsoft's iFilters [3]) and some are not (Oracle's Outside In Technology [5]). Several open source alternatives also exist, such as Apache POI for Microsoft Office documents.

Relying on any one technology, whether free, paid, or open source, is dangerous. Yet, because of the complexity of these file formats, it remains a necessary requirement. By enforcing standards of what metadata and content is to be expected from this extraction technology, we can provide for a more consistent e-discovery experience.

2.3.2 The Need for Open File Formats

An important distinction for these document types is whether the file format is an open standard (email), proprietary yet fully documented (Microsoft Office [4]), or not public information. By specifying the differences between formats, a standard could enforce all data be represented in an open or documented formats. This way, open source solutions, such as Apache Tika [7], can fully participate in e-discovery without fear of reprisal. As a side effect, this could influence holders of closed proprietary formats to open them to the community at large.

One important point, however, deals with the conversion from closed to open formats. As long the standard specifies what content and metadata needs are, the conversion needs to guarantee all data comes across faithfully.

2.3.3 Information in the Cloud

For information residing in the cloud, such as documents in Google Docs, Facebook posts, Twitter updates, etc., determining what is a document can be difficult. Google Docs, for example, saves updates of documents every few seconds. Legally, how can you determine what is a user's intended save point containing a 'coherent' document?

Standardization is even more important here than for known document types—we need to define what a document even means before we can extract metadata and content. Further, all of the metadata we need might not be attached to the content but rather will need to be accessed programmatically.

3. ACKNOWLEDGEMENTS

Thanks to everyone who provided comments and insights on previous drafts of this paper.

4. CONCLUSIONS

In this paper we discussed the need for standards in e-discovery surrounding search query syntax, document encoding, and content extraction. We hope this starts a conversation among e-discovery practioners, search engine vendors, and corporations facing lawsuits with the goal of increasing search quality and consistency during E-Discovery.

5. REFERENCES

[1] Sedona Conference. "The Sedona Conference Homepage" http://www.thesedonaconference.org/, Last accessed 21 April 2011.

[2] Unicode Consortium. "The Unicode Standard", http://www.unicode.org/standard/standard.html. Last accessed 21 April 2011.

[3] Microsoft. "Microsoft Office 2010 Filter Packs". https://www.microsoft.com/downloads/en/details.aspx?FamilyID=5cd4dcd7-d3e6-4970-875e-aba93459fbee, Last accessed 21 April 2011.

[4] Microsoft. "Microsoft Office File Formats", http://msdn.microsoft.com/en-us/library/cc313118%28v=office.12%29.aspx. Last accessed: 21 April 2011.

[5] Oracle. "Oracle Outside In Technology", http://www.oracle.com/us/technologies/embedded/025613.htm. Last accessed: 21 April 2011.

[6] The Apache Foundation. "The Apache POI Project", http://poi.apache.org/. Last accessed: 21 April 2011.

[7] The Apache Foundation. "The Apache Tika Project", http://tika.apache.org/. Last accessed: 21 April 2011.

DESI IV POSITION PAPER

The False Dichotomy of Relevance: The Difficulty of Evaluating the Accuracy of Discovery Review Methods Using Binary Notions of Relevance

BACKGROUND

Manual review of documents by attorneys has been the de facto standard for discovery review in modern litigation. There are many reasons for this, including the inherent authoritativeness of lawyer judgment and presumptions of reliability, consistency, and discerning judgment. In the past couple of decades, growth in the volume of electronic business records has strained the capacity of the legal industry to adapt, while also creating huge burdens in cost and logistics. (Paul and Baron, 2007).

The straightforward business of legal document review has become so expensive and complex that an entire industry has arisen to meet its very particular needs. Continually rising costs and complexity have, in turn, sparked an interest in pursuing alternative means of solving the problem of legal discovery. Classic studies in the field of Information Retrieval which outline the perils and inherent accuracy of manual review processes have found new audiences. (see, e.g. Blair & Maron, 1985) Many newer studies have emerged to support the same proposition, such as the work of the E-Discovery Institute and many who work in the vendor space touting technology-based solutions. Even more recently, cross-pollination from information analytics fields such as Business Intelligence / Business Analytics, Social Networking, and Records Management have begun generating significant "buzz" about how math and technology can solve the problem of human review.

Clients and counsel alike are looking toward different solutions for a very big problem – how to deal with massive amounts of data to find what is important and discharge discovery obligations better and more cost-effectively. The tools available to streamline this job are growing in number and type. Ever more sophisticated search term usage, concept grouping and coding techniques, next generation data visualization techniques, and machine learning approaches are all making inroads into the discovery space. There is ample evidence that the allure of "black box" methods is having an impact on how we believe the problem of large-scale discovery can be resolved.

POSITION

Because math is hard, lawyers have become enamored with notional "process" with its implicit suggestion that there is some metaphysically ideal assembly line approach that can be invoked for each case. All you have to do is make certain tweaks based on case type, complexity, etc. and you will generate a reproducible, defensible product. The approach is analogous to the "lodestar" computation used in assessing the reasonableness of contingency fees in complex cases. This process-focused approach rests on the faulty premise that relevance is an objective, consistently measurable quality, and by extension, that it is susceptible to some objectively measurable endpoint in document review. Deterministic formulas, no matter how sophisticated, can only accounts for discrete variables in the review, such as size, scope, complexity, and the

like. The foundational variable, relevance, is anything but discrete, and without a reproducible, consistent definition of relevance, the input into any formula for review accuracy or success will be unreliable.

The False Dichotomy of Relevance

How do we determine if a document is relevant or not? Disagreement among similarly situated assessors in Information Retrieval studies is a known issue. (Voorhees, 2000). The issue of translating the imperfect, analog world of information to a binary standard of true/false is a difficult one to study. When you compound the confusion by blurring the distinction between relevance, which is something you want, and responsiveness, which is something that may lead to something you want, the difficulty only increases. In practice, this author has participated in side by side testing of learning tools and seen very capable expert trainers develop quite different interpretations of both responsiveness and relevance. Anyone who has been involved in document review understands that where responsiveness or relevance are concerned, reasonable minds can, and often do, disagree. Who is right and who is wrong? Is anyone really right or wrong?

Take note of an actual request by the Federal Trade Commission in antitrust review. It calls for "all documents relating to the company's or any other person's plans relating to any relevant product, including, but not limited to…"

The governing guidance for civil discovery can be found in Federal Rules of Civil Procedure 26(b)(1):

> "Parties may obtain discovery regarding any nonprivileged matter that is relevant to any party's claim or defense – including the existence, description, nature, custody, condition, and location of any documents or other tangible things and the identity and location of persons who know of any discoverable matter… Relevant information need not be admissible at the trial if the discovery appears reasonably calculated to lead to the discovery of admissible evidence."

Very broad requests blended with highly inclusive interpretive guidance give rise to great variability in interpreting both relevance and responsiveness. What is **relevant** gets confused with what is **responsive**, and in both events, a wide range of possible thresholds can be established, depending on who is making the decisions.

As an illustration using the above request, if a reviewer is presented with a document that is a calendar reminder to self concerning a product development meeting that mentions the product by name and a date, but no other information, would it be relevant or responsive? If it mentions other people expected to be in attendance, would that change things? If it also stated the meeting's agenda, what would happen? Depending on the relevant issues of the particular matter, the answers might vary, and this author would disagree that there is any bright line response that covers every use case.

In the bulk of litigation, a large proportion of documents fall into the kind of gray area like the calendar entry example above. There is rarely a hard and fast rule for what is relevant or responsive when context, vernacular, and intent are unknown. Forcing such determinations is a

necessary evil, but distorts conceptions of relevance and responsiveness, particularly when rules and guidance are inferred from prior judgments. The effort of doing so is akin to pushing a round peg through a square hole, and the results are analogous to trying to define obscenity instead of saying "you know it when you see it."

When forcing documents to live in a yes/no world, a marginal yes will be considered the same as an obvious, smoking gun yes for all follow-on evaluations. This creates a problem similar to significant figures calculations in scientific measurement – the incorporation and propagation of uncertainty into further calculations simply yields greater uncertainties. Attempting to adopt objective standards (e.g. F1 measure thresholds) based on a flawed presumption of binary relevance/responsiveness will by extension also be suspect. Comparing different information retrieval and review systems is difficult and often misleading enough without internalizing the uncertainty generated by enforced binary classification of relevance.

Worse yet, the seduction of clean numerical endpoints belies the complexities in deriving them. We would love to say that System A is 90% accurate and System B is 80% accurate, so System A is superior. The truth, however, is that data are different, reviewers are different, assessors are different, and methods of comparing results are different. In the most straightforward matters, there are few documents that are 100% relevant or irrelevant to a given request. Moreover, actual relevance often changes over time and as case issues are defined more narrowly through discovery. After all, if both sides knew everything they needed to know about the case issues at the outset, why bother with discovery?

As recently as the last Sedona annual meeting, there was talk of developing a benchmark F1 measure that could be used as an objectively reasonable baseline for accuracy in a review. This is troubling because even in the most knowledgeable community addressing electronic discovery issues, the notion of an objectively definable standard of relevance/responsiveness is entertained. The legal industry must not succumb to the temptation of easy numbers.[1]

Proposed Solution

Before traveling too far down the road of setting accuracy standards or comparing different review systems, we should question our current conception of notional relevance in legal discovery review and advocate a meaningful, practical approach to benchmarking the accuracy of legal review in the future. We cannot faithfully ascribe *a priori* standards of relevance without the benefit of full knowledge that a real world case will not permit, and we cannot even do a legitimate analysis *ex post facto* unless all stakeholders can agree about what passes muster.

[1] This kind of approach also ignores the fact that statistical measures will not work equally well across different likelihood of responsiveness (e.g. a recall projection for a corpus of 1 million in which 50 docs are truly responsive and 30 are returned would undoubtedly look very different from a projection based on 300,000 found out of 500,000 true responsive). Furthermore, such standard setting does not the fact that different cases call for different standards – a second request "substantial compliance" standard is, in practice, very different from a "leave no stone unturned" standard that one might employ in a criminal matter.

The best we can aim for is to make sure that everyone agrees that what is produced is "good enough." "Good enough" is a fuzzy equation that balances the integrity of the results with the cost of obtaining them, and is evaluated by all concerned parties using their own criteria. Integrity is a utilitarian measure. As a consumer of discovery, a practitioner would want to know that everything that they would be interested in is contained therein. The guidance of the Federal Rules notwithstanding, this does not mean that a recipient of discovery **wants** to know that everything that is arguably responsive is contained in the production corpus, but rather everything that they would deem necessary to flesh out their story and understand / respond to the other side's story is produced. In other words, and at the risk of over-simplification, the consumer of discovery wants some degree of certainty they have received all clearly relevant material. While discovery rules and requests are fashioned to yield the production of documents "tending to lead to the discovery of admissible evidence," this is largely a safety net to ensure no under-production. Analyzing the accuracy of discovery as a function of whether all documents "tending to lead to the discovery of admissible evidence" is a slippery slope. The inquiry quickly turns to determining whether all documents that tend to lead to the discovery of documents that tend to lead to the discovery of potentially admissible evidence, which militates strongly in favor of severe over-production, at considerable cost to both producing and receiving party and also the very system of achieving justice, since it is so fraught with high and avoidable costs.

Relevance within a case is highly volatile, subjective, and particular to that matter. Furthermore, the only parties that care are the ones involved (excluding for the sake of argument those who are interested in broader legal issues at bar). Accordingly, the best way to approach relevance is to adopt some relevance standard that relies on consensus, whether actual, modeled, or imputed. Actual consensus would involve use of representatives of both parties to agree that particular documents are relevant. Modeled consensus would involve using learning systems or predictive algorithms to rank documents according to a descending likelihood of relevance. Imputed consensus would involve the use of a disinterested third party, such as an agreed-upon arbiter or a special master.

The question to be answered by any consensus-based standard should be slightly different than the rather unhelpful "whether this document tends to lead to the discovery of admissible evidence." It should instead focus on actual utility. In terms of defining relevance, perhaps we could articulate the standard as a function of likelihood of being interesting, perhaps "would a recipient of discovery reasonably find this document potentially interesting?" Expressed in the inverse, a non-produced document would be classified as accurately reviewed UNLESS it was clearly interesting. No one really cares about marginally responsive documents, whether they are or are not produced. By extension, we should disregard marginal documents when determining the accuracy of a given review.

As far as applying the standard, there are no objective criteria, so some subjective standard must be applied. This removes the business of assessing review accuracy from the myriad of manufacturing QA/QC processes available, since using objective metrics like load tolerances to measure subjective accuracy is like using word counts to rank the quality of Shakespearean plays. In practice, only the receiving party generally has standing to determine whether or not they are harmed by over or under production, so the most rational approach to determining review quality should begin and end with the use of the receiving party or a reasonable proxy for

them. One possible way of doing this is to assign an internal resource to stand in the shoes of the receiving party and make an independent assessment of samples of production (whether by sampling at different levels of ranked responsiveness, stratified sampling using other dimensions, such as custodian, date, or perhaps search term), and then analyze the results for "clear misses." These clear misses could be converted to a rate of review required to include these (or other metric that demonstrates the diminishing returns associated with pushing the production threshold back), which can then be converted to man-hours and cost to produce such additional documents.

If predictive categorization is being employed, it is also possible to use multiple trainers and then overlay their results. Overlapping results in relevance are a de facto consensus determination, and can be used to ascribe overall responsiveness to a given document. The benefit of this approach is that it also serves a useful QC function.

There are, of course, a number of other possible approaches, but the overriding theme should be that evaluations of effectiveness and accuracy should redraw the lines used to evaluate accuracy, steering away from hard and fast standards and moving toward more consensus-based, matter-specific metrics.

CONCLUSION

Attorneys and the electronic discovery industry should eschew the easy path of arbitrarily derived objective standards to measure quality and accuracy, but at the same time, they cannot expect to develop rigorous, objective rigorous criteria for comparing or evaluating search and review methods. Any evaluation of systems that purport to identify legally relevant or discoverable information rests on a definition of relevance, and relevance is a matter-specific, highly subjective, consensual determination. As a community, we should work toward developing assessment standards that mirror this reality.

Eli Nelson
Cleary, Gottlieb, Steen & Hamilton
2000 Pennsylvania Ave., Washington, DC 20006
(202)974-1874
enelson@cgsh.com

Sampling – The Key to Process Validation

Christopher H. Paskach, Partner, KPMG LLP
Michael J. Carter, Manager, Six Sigma Black Belt, KPMG LLP
May 13, 2011

Abstract
The increasing volume and complexity of electronically stored information and the cost of its review continues to drive the need for development of sophisticated, high-speed processing, indexing and categorization -- or "predictive coding" -- software in response to litigation and regulatory proceedings. Since the majority of these tools rely on sophisticated, proprietary algorithms that are frequently referred to as "black box" technologies, there has been a reluctance to exploit their expected productivity gains for fear that the results they produce may be challenged and rejected as not meeting the required standard of "reasonableness." Effective use of sampling can overcome this concern by demonstrating with a stated level of confidence that the system has produced results at least as consistent and reliable as those obtained by having attorneys review the documents without sophisticated technology support. Through testing, based on statistical sampling, the quality improvements and cost savings promised by "predictive coding" technology can be realized.

Current State
Determining the reasonableness of a document search and review process has been based on whether there was sufficient "attorney review" of the documents to ensure that the results will be reliable. While "attorney review" has been accepted as the "gold standard" for the adequacy of a document review process, the consistency and reliability of the results produced by the attorneys has rarely been questioned or tested. The presumed effectiveness of "attorney review" is generally accepted to meet the reasonableness standard so that sophisticated sampling and testing of the attorney review is rarely performed. The sheer volumes and unforgiving production deadlines of today's e-discovery efforts demand ever increasing review capacity and throughputs. Simply scaling up the review process with more people to handle these demands is clearly at odds with cost control initiatives that are of utmost importance to corporate law departments.

New Technologies
Recent technology development efforts have focused primarily on helping review teams manage the efficiency and cost of large scale reviews. Of particular interest are the tools that help categorize and cluster documents based on document content, or identifying near-duplicate documents and grouping them for review. These "predictive coding" tools can help reviewers speed through non-responsive or similar sets of documents by bulk tagging and more quickly isolating relevant material that has to be more carefully reviewed for privilege before being produced. The very latest technologies aim to automate the review process by minimizing the need for human reviewers in a first-pass review for relevance. But regardless of where an organization falls on the automation continuum in its adoption of technology -- from traditional linear review to concept-based clustering, leveraging technology or human review -- the goal of a faster, more consistent, more predictable and less costly review requires

more than basic efficiency gains. A cost-effective document review project requires more sophisticated technology and proven Quality Control (QC) processes to demonstrate its effectiveness.

Technology vs. Human Review

While an argument for cost effectiveness of technology-based processes has been largely established, the consistency of the results "versus" human review remains a topic of ongoing discussion. For many, the validation of review quality is often subordinate to the review itself and consists of informal or casual observations that lack the scientific rigor and quantifiable measures necessary to defend the quality proc*e*ss. More sophisticated quality methods that rely on sampling can provide the much needed assurance that the results are at least as good as human review and when used appropriately can result in significantly improved consistency and productivity.

Over the past three years, KPMG has conducted four test projects that compared the results of an "attorney review" process with results obtained by reprocessing the same document collection with a predictive-coding software tool. The tool used in these tests uses a series of randomly selected sample batches of 40 documents that are reviewed by a subject matter expert (SME) to train the software. Based on the SME's decisions on the training batches, the software calculates the relevance of the remaining documents in the collection. In all four test cases, the software was more consistent in categorizing documents than were the human review teams.

Although the software produced more consistent results than the review attorneys, the proprietary algorithm used to produce the relevance ranking is not publicly available. However, the results it produces can be effectively tested with sampling to determine the efficacy of the automated relevance ranking process.

Assuring Process Quality

Assuring process capability, explicitly or implicitly, is a requirement for defensibility. Having a defendable, and therefore accepted, process is a matter of sound design, transparency and predictable results. Process sampling delivers all three requirements. Sampling is a well proven, scientifically rigorous method that can give the Project Manager much needed flexibility to demonstrate effectively the quality of the review process. Carefully selecting a sample, and from it inferring the condition of the larger population with high confidence in the reliability of the inference, is a powerful tool with tremendous eDiscovery utility. The process of establishing review QC using statistical sampling enables the review team to determine appropriate sample size, quantify the process risks, and determine process acceptance and rejection criteria. Then, should questions arise concerning the quality of the results, a meaningful discussion of the QC methodology can take place without the need to explain, justify or alter unproven judgmental QC practices.

Objections to Statistical Sampling

If statistical sampling can provide all of these benefits to QC in discovery review, why isn't it more widely used? There are several possible reasons, including a lack of familiarity with the method or its perceived complexity and the anticipated time investment required to understand and achieve proficiency in it. Another concern may be that a small error found in sampling could render the entire review results unacceptable. Likewise, in the discovery review process there is no clear legal precedent that confirms

the acceptability of statistical sampling methods for eDiscovery. Whatever the reasons, although sampling is widely accepted as a basic QC methodology in numerous other product and service industries to manage and quantify quality risk, it has not been widely adopted in eDiscovery review projects.

Overcoming the Objections

How can the issues that prevent wider use of statistical sampling be addressed? Overcoming the lack of familiarity with sampling can be addressed through training and the use of experts. Involving those who understand QC sampling in the process of eDiscovery can be a very effective approach to achieving the benefits and overcoming project managers' unfamiliarity. These sampling experts can assist with data stratification, determining sample sizes and calculating confidence levels for statistical inferences. One objection to this approach would be the added cost of these sampling experts. This can be addressed with a straight-forward cost-benefit calculation comparing the cost of the experts to the avoided costs of more extensive testing with non-statistical approaches. Another objection would be the risk of the supervising attorneys not being sufficiently knowledgeable to assess the quality of the sampling experts' work. This can be addressed through careful questioning and review of the experts' approach and results.

Another option to support using statistical sampling would be to programmatically integrate generally accepted QC sampling methods into widely-used eDiscovery applications. Carefully designed user interfaces for selecting samples, testing them and reporting the results could guide users through the sampling process, thereby minimizing, if not eliminating, most common sampling mistakes. Increased consistency, repeatability and reproducibility of the QC process would result.

Additionally, the sampling methodology could include periodic batch sampling throughout the review process with a mechanism for dealing with review error as soon as it is detected to reduce the need to re-perform a significant portion of the review process. Likewise, sampling error could be addressed with a set of tools that would enable sample results to be adjusted and reinterpreted in light of sampling error to reduce the risk of having to significantly expand the sample or restart the sampling process.

The final objection regarding a lack of a clear legal precedent is likely to be addressed soon by the courts, which are becoming increasingly aware of the benefits of statistical sampling in dealing with the challenges posed by very large populations of documents. Without clear legal precedent there is some additional risk to applying new technologies and relying on statistical sampling to demonstrate their efficacy. However, the benefits in terms of quality and cost of QC sampling the results from these new technologies can more than offset these risks until the legal precedents supporting their use are clearly established.

Note: The preceding commentary relates solely to process control sampling as applied in the performance of document review in connection with electronic discovery and is NOT a commentary on the maturity of sampling techniques relative to financial statement auditing.

Process Evaluation in eDiscovery as Awareness of Alternatives

Jeremy Pickens, John Tredennick, Bruce Kiefer
Catalyst Repository Systems
1860 Blake Street, 7th Floor
Denver, Colorado
303.824.0900
{jpickens, jtredennick, bkiefer}@catalystsecure.com

ABSTRACT
With a growing willingness in the legal community to accept various forms of algorithmic augmentation of the eDiscovery process, better understanding of the quality of these machine-enhanced approaches is needed. Our view in this position paper is that one of the more important ways to understand quality is not in terms of absolute metrics on the algorithm, but in terms of an understanding of the effectiveness of the alternative choices a user could have made while interacting with the system. The user of an eDiscovery platform needs to know not only how well an information seeking process is running, but how well the alternatives to that process could have run.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – Search process

General Terms
Measurement, Experimentation, Standardization.

Keywords
Iterative Information Seeking, Interactive Information Seeking, eDiscovery, Process Evaluation

1. INTRODUCTION
Unlike traditional ad hoc search (such as web search) in which the information seeking process is typically single-shot, eDiscovery has both the potential and the necessity to be iterative. Information needs in eDiscovery-oriented information seeking are changing and ongoing, and often cannot be met in a single round of interaction or by a single query. Evaluation of eDiscovery platform quality must take this into account.

There are many metrics for single-shot, non-interactive retrieval, such as precision, recall, mean average precision, and PRES [1]. Our goal is not to propose a new single-shot metric. Instead, we declare that what is needed is an approach in which any or all of these metrics are used in an interactive context.

Furthermore, we take a user-centric view in that we are not concerned with comparison between eDiscovery platforms but instead are concerned with helping the user understand where he or she is within a larger information seeking task on a single platform. A quality process should be one in which the user is able to both (1) affect system behavior by making conscious choices, and (2) explicitly obtain an understanding of the consequences of those choices, so as to adapt and make better choices in the future.

2. THE "WHAT IF" OF EDISCOVERY

2.1 Choices
Interactive information seeking in general and eDiscovery in particular are characterized by choices. Even with machine augmentation of the search process there is still a human in the loop, considering alternatives and making decisions. Examples of choices, not all of which are independent of each other, include:

1. Does one continue traversing the results list for an existing query, or does one issue a new query instead
2. If complete queries are offered as suggestions, which of the alternatives does one pick?
3. If individual terms are offered as query expansion options, which of the alternatives does one pick, and when does one stop adding additional terms?
4. If the collection is clustered in some manner, which cluster does one choose to examine, and when does one stop examining that cluster?
5. If multiple sources (e.g. custodians) or document types (e.g. PDF, PPT, Word, email) are available, how does one choose which sources or types to pay the most attention to?
6. When the document volumes go beyond what is feasible to review, how do you determine when to stop reviewing?
7. At what point do you produce documents which haven't been personally reviewed.

2.2 Consequences
In the previous section we outlined a few examples of the types of choices that an information seeker has to make. Each of those choices has consequences. The choice to dedicate time and resources investigating information coming from one custodian means that less time and fewer resources will be dedicated to a different custodian. More time spent traversing the result set of one query means less time spent on the results of a different query, or perhaps fewer queries executed overall. Adding some terms to an existing query (during query expansion) means not adding others. Deciding that a particular point would be a good one at which to stop reviewing, and then continuing to review anyway might yield diverging expectations as new pockets or rich veins of information are discovered.

In order to understand the quality of a search process, knowing the effectiveness of such choices are not enough. A user has to be able to come to know and understand the opportunity costs of the choices not taken. Does an eDiscovery platform make it possible for a user to understand the consequences of his or her choices?

Does the system give a user a working awareness of the alternatives? Is it possible for the user to return to a previous choice at a later point in time and obtain feedback on the question of "what if" that path had been chosen? A quality search process should be able to answer, or at least give insight into, these questions.

3. PRINCIPLES AND EXAMPLES

Giving an information seeker an awareness of alternatives is not an approach tied to any one particular algorithmically-enhanced methodology. The manner in which a machine (algorithm) learns from the human and applies that learning to the improvement of future choices is a separate issue from whether or not the user is able to garner insight into the efficacy of alternative choices. Granted, some algorithmic approaches might be more penetrable, more conducive to proffering the needed awareness. But the feedback on choices taken versus not taken are going to depend heavily on the nature of the choices themselves.

That said, we offer a few principles which might aide in the design of consequence-aware systems:

1. If there is overlap between the multiple choices (i.e. if the consequences of certain choices are not mutually exclusive) then information garnered while following one choice could be used to make inferences about another choice.
2. If there is overlap between the consequences (results) of a single choice, then the efficacy of that choice can be more quickly assessed by examining fewer, perhaps more "canonical" results.

For example, a clustering algorithm might not partition a set of documents, but instead place a few of the same documents in multiple clusters. Or the same (duplicate or near-duplicate) documents might be found in the collections from more than one custodians. Or two different query expansion term choices (e.g. "bees" and "apiary") might retrieve many of the same documents. In such cases, judgments (coding) on these shared documents can be used to assess multiple choices. Naturally the assessment is done within the context of whatever metric is most important to the user, whether that metric is precision, recall, or something else entirely. But the principle of using overlap to estimate and make inferences on that metric remains.

The way in which this could be made to work would be to implement a process-monitoring subsystem that keeps track of choices both taken and not taken, and then uses information such as the ongoing manual coding of responsiveness and privilege to assess the validity of those choices. The differential between expectation at one point in time and reality at a future point in time should yield more insight into the information seeking eDiscovery process than just knowing the precision or recall effectiveness at any given point in time.

4. ISSUES

The largest issue that needs to be resolved for alternative-aware approaches is that of ever-expanding choice. At every round of interaction, at every point in the human-machine information seeking loop at which the human has the ability to make a choice, a number of options become available. Every choice then gives rise to another set of choices, in an exponentially-branching set of alternatives. Naturally this exponential set needs to be pruned into a manageable set of the most realistic, or possibly the most diverse, set of alternatives.

The consequences of every possible choice or path not taken probably do not to be tracked and monitored; a subset should be fine. However, there needs to be enough awareness of alternatives that the user can get an overall sense of how well he or she is doing, and how much progress is or is not being made with respect to the other choices that were available at various stages. The user needs to be able to get a sense of how well a choice at one point in time matches reality as the consequences of that and other, hypothetically-followed choices become clearer at later points in time.

5. SUMMARY

Information retrieval has a long history of using user interaction (e.g. in the form of relevance feedback and query expansion, for example) to improve the information seeking process in an iterative manner. User behavior alters the algorithm. However, it is also true that the algorithm alters the user. The more choices a user makes, the more potential exists that some of these choices are sub-optimal. Therefore, awareness of alternative choices are needed to help the user orient himself inside of complex information seeking tasks such as in eDiscovery. This paper proposes an approach to the evaluation of quality not in terms of system comparison, but in terms of alternative, path-not-taken choice comparison and awareness.

6. REFERENCES

[1] Magdy, Walid and Jones, Gareth. *In the Proceedings of the 33rd Annual SIGIR Conference.* PRES: A Score Metric for Evaluating Recall-Oriented Information Retrieval Applications. Geneva, Switzerland. August 2010.

Discovery of Related Terms in a corpus using Reflective Random Indexing

Venkat Rangan

Clearwell Systems, Inc.

venkat.rangan@clearwellsystems.com

ABSTRACT

A significant challenge in electronic discovery is the ability to retrieve relevant documents from a corpus of unstructured text containing emails and other written forms of human-to-human communications. For such tasks, recall suffers greatly since it is difficult to anticipate all variations of a traditional keyword search that an individual may employ to describe an event, entity or item of interest. In these situations, being able to automatically identify conceptually related terms, with the goal of augmenting an initial search, has significant value. We describe a methodology that identifies related terms using a novel approach that utilizes Reflective Random Indexing and present parameters that impact its effectiveness in addressing information retrieval needs for the TREC 2010 Enron corpus.

1. Introduction

This paper examines reflective random indexing as a way to automatically identify terms that co-occur in a corpus, with a view to offering the co-occurring terms as potential candidates for query expansion. Expanding a user's query with related terms either by interactive query expansion [1, 5] or by automatic query expansion [2] is an effective way to improve search recall. While several automatic query expansion techniques exist, they rely on usage of a linguistic aid such as thesaurus [3] or concept-based interactive query expansion [4]. Also, methods such as ad-hoc or blind relevance feedback techniques rely on an initial keyword search producing a top-n results which can then be used for query expansion.

In contrast, we explored building a semantic space using Reflective Random Indexing [6, 7] and using the semantic space as a way to identify related terms. This would then form the basis for either an interactive query expansion or an automatic query expansion phase.

Semantic space model utilizing reflective random indexing has several advantages compared to other models of building such spaces. In particular, for the specific workflows typically seen in electronic discovery context, this method offers a very practical solution.

2. Problem Description

Electronic discovery almost always involves searching for relevant and/or responsive documents. Given the importance of e-discovery search, it is imperative that the best technologies are applied for the task. Keyword based search has been the bread and butter method of searching, but its limitations have been well understood and documented in a seminal study by Blair & Moran [8]. At its most basic level, concept search technologies are designed to overcome some limitations of keyword search.

When applied to document discovery, traditional Boolean keyword search often results in sets of documents that include non-relevant items (false positives) or that exclude relevant terms (false negatives). This is primarily due to the effects of synonymy (different words with similar meanings) or polysemy (same word with multiple meanings). For polysemes, an important characteristic requirement is that they share the same etymology but their usage has evolved it into different meanings. In addition, there are also situations where words that do not share the same etymology have different meanings (e.g., river bank vs. financial bank), in which case they are classified as homonyms.

In addition to the above word forms, unstructured text content, and especially written text in emails and instant messages contain user-created code words, proper name equivalents, contextually defined substitutes, and prepositional references etc., that mask the document from being indentified using Boolean keyword search. Even simple misspellings, typos and OCR scanning errors can make it difficult to locate relevant documents.

Also common is an inherent desire of speakers to use a language that is most suited from the perspective of the speaker. The Blair Moran study illustrates this using an event which the victim's side called the event in question an "accident" or a "disaster" while the plaintiff's side called it an "event", "situation", "incident", "problem", "difficulty", etc. The combination of human emotion, language variation, and assumed context makes the challenge of retrieving these documents purely on the basis of Boolean keyword searches an inadequate approach.

Concept based searching is a very different type of search when compared to Boolean keyword search. The input to concept searching is one or more words that allow the investigator or user to express a concept. The search system is then responsible for identifying other documents that belong to the same concept. All concept searching technologies attempt to retrieve documents that belong to a concept (reduce false negatives and improve recall) while at the same time not retrieve irrelevant documents (reduce false positives and increase precision).

3. Concept Search approaches

Concept search, as applied to electronic discovery, is a search using meaning or semantics. While it is very intuitive in evoking a human reaction, expressing meaning as input to a system and applying that as a search that retrieves relevant documents is something that requires a formal model. Technologies that attempt to do this formalize both the input request and the model of storing and retrieving potentially relevant documents in a

mathematical form. There are several technologies available for such treatment, with two broad overall approaches: unsupervised learning and supervised learning. We examine these briefly in the following sections.

3.1 Unsupervised learning

These systems convert input text into a semantic model, typically by employing a mathematical analysis technique over a representation called vector space model. This model captures a statistical signature of a document through its terms and their occurrences. A matrix derived from the corpus is then analyzed using a Matrix decomposition technique.

The system is unsupervised in the sense that it does not require a training set where data is pre-classified into concepts or topics. Also, such systems do not use ontology or any classification hierarchy and rely purely on the statistical patterns of terms in documents.

These systems derive their semantics through a representation of co-occurrence of terms. A primary consideration is maintaining this co-occurrence in a form that reduces impact of noise terms while capturing the essential elements of a document. For example, a document about an automobile launch may contain terms about automobiles, their marketing activity, public relations etc., but may have a few terms related to the month, location and attendees, along with frequently occurring terms such as pronouns and prepositions. Such terms do not define the concept automobile, so their impact in the definition must be reduced. To achieve such end result, unsupervised learning systems represent the matrix of document-terms and perform a mathematical transformation called dimensionality reduction. We examine these techniques in greater detail in subsequent sections.

3.2 Supervised learning

In the supervised learning model, an entirely different approach is taken. A main requirement in this model is supplying a previously established collection of documents that constitutes a training set. The training set contains several examples of documents belonging to specific concepts. The learning algorithm analyzes these documents and builds a model, which can then be applied to other documents to see if they belong to one of the several concepts that is present in the original training set. Thus, concept searching task becomes a concept learning task.

It is a machine learning task with one of the following techniques.

a) Decision Trees
b) Naïve Bayesian Classifier
c) Support Vector Machines

While supervised learning is an effective approach during document review, its usage in the context of searching has significant limitations. In many situations, a training set that covers all possible outcomes is unavailable and it is difficult to locate exemplar documents. Also, when the number of outcomes is very large and unknown, such methods are known to produce inferior results.

For further discussion, we focus on the unsupervised models, as they are more relevant for the particular use cases of concept search.

3.3 Unsupervised Classification Explored

As noted earlier, concept searching techniques are most applicable when they can reveal semantic meanings of a corpus without a supervised learning phase. To further characterize this technology, we examine various mathematical methods that are available.

3.4 Latent Semantic Indexing

Latent Semantic Indexing is one of the most well-known approaches to semantic evaluation of documents. This was first advanced in Bell Labs (1985), and later advanced by Susan Dumais and Landauer and further developed by many information retrieval researchers. The essence of the approach is to build a complete term-document matrix, which captures all the documents and the words present in each document. Typical representation is to build an N x M matrix where the N rows are the documents, and M columns are the terms in the corpus. Each cell in this matrix represents the frequency of occurrence of the term at the “column” in the document “row”.

Such a matrix is often very large – document collections in the millions and terms reaching tens of millions are not uncommon. Once such a matrix is built, mathematical technique known as Singular Value Decomposition (SVD) reduces the dimensionality of the matrix into a smaller size. This process reduces the size of the matrix and captures the essence of each document by the most important terms that co-occur in a document. In the process, the dimensionally reduced space represents the “concepts” that reflect the conceptual contexts in which the terms appear.

3.5 Principal Component Analysis

This method is very similar to latent semantic analysis in that a set of highly correlated artifacts of words and documents in which they appear, is translated into a combination of the smallest set of uncorrelated factors. These factors are the principal items of interest in defining the documents, and are determined using a singular value decomposition (SVD) technique. The mathematical treatment, application and results are similar to Latent Semantic Indexing.

A variation on this, called independent component analysis is a technique that works well with data of limited variability. However, in the context of electronic discovery documents where data varies widely, this results in poor performance.

3.6 Non-negative matrix factorization

Non-negative matrix factorization (NMF) is another technique most useful for classification and text clustering where a large collection of documents are forced into a small set of clusters. NMF constructs a document-term matrix similar to LSA and includes the word frequency of each term. This is factored into a term-feature and feature-document matrix, with the features automatically derived from the document collection. The process also constructs data clusters of related documents as part of the mathematical reduction. An example of this research is available at [2] which takes the Enron email corpus and classifies the data using NMF into 50 clusters.

3.7 Latent Dirichlet Allocation

Latent Dirichlet Allocation is a technique that combines elements of Bayesian learning and probabilistic latent semantic indexing. In this sense, it relies on a subset of documents pre-classified into a training set, and unclassified documents are classified into

concepts based on a combination of models from the training set [15].

3.8 Comparison of the above technologies

Although theoretically attractive and experimentally successful, word space models are plagued with efficiency and scalability problems. This is especially true when the models are faced with real-world applications and large scale data sets. The source of these problems is the high dimensionality of the context vectors, which is a direct function of the size of the data. If we use document-based co-occurrences, the dimensionality equals the number of documents in the collection, and if we use word-based co-occurrences, the dimensionality equals the vocabulary, which tends to be even bigger than the number of documents. This means that the co-occurrence matrix will soon become computationally intractable when the vocabulary and the document collections grow.

Nearly all the technologies build a word space by building a word-document matrix with each row representing a document and column representing a word. Each cell in such a matrix represents the frequency of occurrence of the word in that document. All these technologies suffer from a memory space challenge, as these matrices grow to very large sizes. Although many cells are sparse, the initial matrix is so large that it is not possible to accommodate the computational needs of large electronic discovery collections. Any attempt to reduce this size to a manageable size is likely to inadvertently drop potentially responsive documents.

Another problem with all of these methods is that they require the entire semantic space to be constructed ahead of time, and are unable to accommodate new data that would be brought in for analysis. In most electronic discovery situations, it is routine that some part of the data is brought in as a first loading batch, and once review is started, additional batches are processed.

4. Reflective Random Indexing

Reflective random indexing (RRI) [6, 7, 11] is a new breed of algorithms that has the potential to overcome the scalability and workflow limitations of other methods. RRI builds a semantic space that incorporates a concise description of term-document co-occurrences. The basic idea of the RRI and the semantic vector space model is to achieve the same dimensionality reduction espoused by latent semantic indexing, without requiring the mathematically complex and intensive singular value decomposition and related matrix methods. RRI builds a set of semantic vectors, in one of several variations – term-term, term-document and term-locality. For this study, we built an RRI space using term-document projections, with a set of term vectors and a set of document vectors. These vectors are built using a scan of the document and term space with several data normalization steps.

The algorithm offers many parameters for controlling the generation of semantic space to suit the needs of specific accuracy and performance targets. In the following sections, we examine the elements of this algorithm, its characteristics and various parameters that govern the outcome of the algorithm.

4.1 Semantic Space Construction

As noted earlier, the core technology is the construction of semantic space. A primary characteristic of the semantic space is a term-document matrix. Each row in this matrix represents all documents a term appears in. Each column in that matrix represents all terms a document contains. Such a representation is an initial formulation of the problem for vector-space models. Semantic relatedness is expressed in the connectedness of each matrix cell. Two documents that share the same set of terms are connected through a direct connection. It is also possible for two documents to be connected using an indirect reference.

In most cases, term-document matrix is a very sparse matrix and can grow to very large sizes for most document analysis cases. Dimensionality reduction reduces the sparse matrix into a manageable size. This achieves two purposes. First, it enables large cases to be processed in currently available computing platforms. Second, and more importantly, it captures the semantic relatedness through a mathematical model.

The RRI algorithm begins by assigning a vector of a certain dimension to each document in the corpus. These assignments are chosen essentially at random. For example, the diagram below has assigned a five-dimensional vector to each document, with specific randomly chosen numbers at each position. These numbers are not important – just selecting a unique pattern for each document is sufficient.

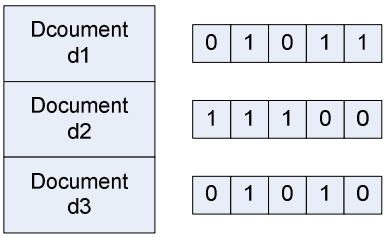


Figure 1: Document Vectors

From document vectors, we construct term vectors by iterating through all terms in the corpus, and for each term, we identify the documents that term appears in. In cases where the term appears multiple times in the same document, that term is given a higher weight by using its term frequency.

$$t_{i,j} = \sum_{k=0}^{k} n_k d_{k,j}$$

Each term k's frequency in the document n_k weighs in for each document vector's position. Thus, this operation projects all the documents that a term appears in, and condenses it into the dimensions allocated for that term. As is evident, this operation is a fast scan of all terms and their document positions. Using Lucene API *TermEnum* and *TermDocs*, a collection of term vectors can be derived very easily.

Once the term vectors are computed, these term vectors are projected back on to document vectors. We start afresh with a new set of document vectors, where each vector is a sum of the term vectors for all the terms that appear in that document. Once again, this operation is merely an addition of floating point numbers of each term vector, adjusting for its term frequency in that document. A single sweep of document vectors to term vector projection followed by term vectors to document vector constitutes a training cycle. Depending on needs of accuracy in the construction of semantic vectors, one may choose to run the

training cycle multiple times. Upon completion of the configured number of training cycles, document and term vector spaces are persisted in a form that enables fast searching of documents during early data exploration, search, and document review.

It is evident that by constructing the semantic vector space, the output space captures the essential co-occurrence patterns embodied in the corpus. Each term vector represents a condensed version all the documents the term appears in, and each document vector captures a summary of the significant terms present in the document. Together, the collection of vectors represents the semantic nature of related terms and documents.

Once a semantic space is constructed, a search for related terms of a given query term is merely a task of locating the nearest neighbors of the term. Identifying such terms involves using the query vector to retrieve other terms in the term vector stores which are closest to it by cosine measurement. Retrieving matching documents for a query term is by identifying the closest documents to the query term's vector in document vector space, again by way of cosine similarity.

An important consideration for searching vector spaces is the performance of locating documents that are cosine-similar, without requiring a complete scan of the vector space. To facilitate this, the semantic vector space is organized in the form of clusters, with sets of the closest vectors characterized by both its centroid and the Euclidean distance of the farthest data point in the cluster. These are then used to perform a directed search eliminating the examination of a large number of clusters.

4.2 Benefits of Semantic Vector Space

From the study the semantic vector space algorithm, one can immediately notice the simplicity in realizing the semantic space. A linear scan of terms, followed by a scan of documents is sufficient to build a vector space. This simplicity in construction offers the following benefits.

a) In contrast to LSA and other dimensionality reduction techniques the semantic space construction requires much less memory and CPU resources. This is primarily because matrix operations such as singular value decomposition (SVD) are computationally intensive, and requires both the initial term-document matrix and intermediate matrices to be manipulated in memory. In contrast, semantic vectors can be built for a portion of the term space, with a portion of the index. It is also possible to scale the solution simply by employing persistence to disk at appropriate batching levels, thus scaling to unlimited term and document collections.

b) The semantic vector space building problem is more easily parallelizable and distributable across multiple systems. This allows parallel computation of the space, allowing for a distributed algorithm to work on multiple term-document spaces simultaneously. This can dramatically increase the availability of concept search capabilities to very large matters, and within time constraints that are typically associated with large electronic discovery projects..

c) Semantic space can be built incrementally, as new batches of data are received, without having to build the entire space from scratch. This is a very common scenario in electronic discovery, as an initial batch of document review needs to proceed before all batches are collected. It is also fairly common for the scope of electronic discovery to increase after early case assessment.

d) Semantic space can be tuned using parameter selection such as dimension selection, similarity function selection and selection of term-term vs. term-document projections. These capabilities allow electronic discovery project teams to weigh the costs of computational resources against the scope of documents to be retrieved by the search. If a matter requires a very narrow interpretation of relevance, the concept search algorithm can be tuned and iterated rapidly.

Like other statistical methods, semantic spaces retain their ability to work with a corpus containing documents from multiple languages, multiple data types and encoding types etc., which is a key requirement for e-discovery. This is because the system does not rely on linguistic priming or linguistic rules for its operation.

5. Performance Analysis

Resource requirements for building a semantic vector space is an important consideration. We evaluated the time and space complexity of semantic space algorithms as a function of corpus size, both from the initial construction phase and for follow-on search and retrievals.

Performance measurements for both aspects are characterized for four different corpora, as indicated below.

Corpus	**Reuters Collection**	**EDRM Enron**	**TREC Tobacco Corpus**
PST Files	-	171	-
No. of Emails	-	428072	-
No. of Attachments	21578	305508	6,270,345
No. of Term Vectors (email)	-	251110	-
No. of Document Vectors (email)	-	402607	-
No. of Term Vectors (attachments)	63210	189911	3,276,880
No. of Doc Vectors (attachments)	21578	305508	6,134,210
No. of Clusters (email)	-	3996	-
No. of Clusters (attachments)	134	2856	210,789

Table 1: Data Corpus and Semantic Vectors

As can be observed, term vectors and document vectors vary based on the characteristics of the data. While the number of

document vectors closely tracks the number of documents, the number of term vectors grows more slowly. This is the case even for OCR-error prone ESI collections, where the term vector growth moderated as new documents were added to the corpus.

5.1 Performance of semantic space building phase

Space complexity of the semantic space model is linear with respect to the input size. Also, our implementation partitions the problem across certain term boundaries and persists the term and document vectors for increased scalability. The algorithm requires memory space for tracking one million term and document vectors, which is about 2GB, for a semantic vector dimension of 200.

Time for semantic space construction is linear on the number of terms and documents. For very large corpus, the space construction requires periodic persistence of partially constructed term and document vectors and their clusters. A typical configuration persist term vectors for each million terms, and documents at each million documents. As an example, the TREC tobacco corpus would require 4 term sub-space constructions, with six document partitions, yielding 24 data persistence invocations. If we consider the number of training cycles, each training cycle repeats the same processes. As an example, the TREC tobacco corpus with two training cycles involves 48 persistence invocations. For a corpus of this size, persistence adds about 30 seconds for each invocation.

Performance Item	Vector Construction (minutes)	Cluster Construction (minutes)
Reuters-21578 dataset	1	1
EDRM Enron dataset	40	15
TREC Tobacco Corpus	490	380

Table 2: Time for space construction, two training cycles (default)

These measurements were taken on commodity Dell PowerEdge R710 system, with two Quad Xeon 5500 processors at 2.1GHz CPU and 32GB amount of memory.

5.2 Performance of exploration and search

Retrieval time for a concept search and time for building semantic space exploration are also characterized for various corpus sizes and complexity of queries. To facilitate a fast access to term and document vectors, our implementation has employed a purpose-built object store. The object store offers the following.

a) Predictable and consistent access to a term or document semantic vector. Given a term or document, the object store provides random access and retrieval to its semantic vector within 10 to 30 milliseconds.

b) Predictable and consistent access to all nearest neighbors (using cosine similarity and Euclidean distance measures) of a term and document vector. The object store has built-in hierarchical k-means based clustering. The search algorithm implements a cluster exploration technique that algorithmically chooses the smallest number of clusters to examine for distance comparisons. A cluster of 1000 entries is typically examined in 100 milliseconds or less.

Given the above object store and retrieval paths, retrieval times for searches range from 2 seconds to 10 seconds, depending on large part, on the number of nearest neighbors of a term, the number of document vectors to retrieve and on the size of the corpus.

The following table illustrates observed performance for the Enron corpus, using the cluster-directed search described above.

Term vector search	Average	Stdev
Clusters Examined	417.84	274.72
Clusters Skipped	1001.25	478.98
Terms Compared	24830.38	16079.72
Terms Matched	21510.29	15930.2
Total Cluster Read Time (ms)	129.39	88.23
Total Cluster Read Count	417.84	274.72
Average Cluster Read Time (ms)	0.29	0.18
Total Search Time (ms)	274.56	187.27

Table 3: Search Performance Measurements

As is apparent from the above time measurements as well as number of clusters examined and skipped, identifying related terms can be offered to users with predictability and consistency, thereby making it possible for its usage as an interactive, exploratory tool during early data analysis, culling, analysis and review phases of electronic discovery.

6. Search Effectiveness

An important analysis is to evaluate the effectiveness of retrieval of related terms from the perspective of the search meeting the information retrieval needs of the e-discovery investigator. We begin by analyzing qualitative feel for search results by examining the related terms and by identifying the relevance of these terms. We then analyze search effectiveness using the standard measures, Precision and Recall. We also examine search effectiveness using Discounted Cumulative Gain (DCG).

6.1 Qualitative Assessment

To obtain a qualitative assessment, we consider the related terms retrieved and examine its nearness measurement, and validate the closest top terms. The nearness measure we use for this analysis is a cosine measure of the initial query vector when compared with the reported result. It is a well-understood measure of judgment of quality in that a cosine measure reflects the alignment of the two vectors, and closeness to the highest value of cosine, which is 1.0, means perfect alignment.

Table 4 shows alignment measures for two concept query terms for the EDRM Enron Dataset [12].

It is quite clear that several of the related terms are in fact logically related. In cases where the relationship is suspect, it is indeed the case that co-occurrence is properly represented. E.g., the term *offshore* and *mainly* appear in enough documents together to make it to the top 20 related terms. Similarly, we have *offshore* and *foreign* co-occur to define the concept of *offshore* on the basis of the identified related terms.

Query: drilling		**Query: offshore**	
Related Term	**Similarity**	**Related Term**	**Similarity**
refuge	0.15213	interests	0.13212
Arctic	0.12295	foreign	0.13207
wildlife	0.12229	securing	0.12597
exploration	0.11902	viable	0.12422
Rigs	0.11172	involves	0.12345
Rig	0.11079	associated	0.12320
supplies	0.11032	mainly	0.12266
Oil	0.11017	principle	0.12248
refineries	0.10943	based	0.12241
Environmentalists	0.10933	achieved	0.12220

Table 4: Illustration of two query terms and their term neighbors

We can further establish the validity of our qualitative assessment using individual document pairs and their document co-occurrence patterns. As an example, Table 5 shows cosine similarity, the number of documents the two terms appear in and the common set of documents both terms appear in, again in the EDRM Enron Dataset.

Term1	**Term2**	**Cosine**	**Docs1**	**Docs2**	**CDocs**
offshore	drilling	0.2825	1685	1348	572
governor	Davis	0.3669	2023	2877	943
brownout	power	0.0431	13	30686	13
brownout	ziemianek	0.5971	13	2	1

Table 5: Cosine similarity comparison for select terms from EDRM Enron corpus

An observation from the above data is that when the two terms compared appear in large number of documents with large overlap, the similarity is greater. In contrast, if one term is dominant in its presence in a large number of documents, and the other term is not, the presence of the two terms in all the common documents (*brownout* and *power*), the similarity is lower. Also noteworthy is if two terms are common in every document and the documents each appears in are small number (*brownout* and *ziemianek*) the similarity measure is significantly higher.

6.2 Precision and Recall Measures

Precision and recall are two widely used metrics for evaluating the correctness of a search algorithm [8]. Precision refers to the ratio of relevant results compared to the full retrieved set, and represents the number of false positives in the result. Recall on the other hand, measures the ratio of relevant results compared to the number of relevant results actually present in the collection, i.e. the number of false negatives. Usually, recall is a harder measure to determine since it would require reviewing the entire collection for identifying all the relevant items, and sample-based estimation is a substitute.

For our purposes, two critical information retrieval needs should be evaluated.

a) The ability of the system to satisfy information retrieval needs for the related concept terms.
b) The ability of the system to provide the same for documents in a concept.

We evaluated both for several specific searches using the EDRM Enron dataset, and we present our results below.

6.3 Precision and Recall for Related Concept Terms

Note that Precision and Recall are defined for related concept terms using a combination of automated procedures and manual assessment. As an example, we supplied a list of queries and their related concept terms and asked human reviewers to rate each related term result as either strongly correlated or related to the initial query, or if it is not related. This gives us an indication of precision for our results, for a given cutoff point. Evaluating recall is harder, but we utilized a combination of sampling methodology and a deeper probe into related term result. As an example of this, we evaluated precision for a cutoff at 20 terms and recall by examining 200 terms and constructing relevance graphs.

6.4 Impact of dimensions

Given that the semantic vector space performs a dimensionality reduction, we were interested in understanding the impact of dimension choice for our semantic vectors. For the default implementation, we have a vector dimension of 200, which means that each term and document has a vector of 200 floating point numbers.

To study this, we performed a study of precision and recall for the EDRM Enron dataset and tracked the precision-recall graph for four choices of dimensions. The results are indicated in Figure 2 below.

As can be observed, we did not gain significant improvement on precision and recall characteristics with a higher choice of dimension. However, for a large corpus, we expect that precision-recall graph would indicate a significantly steeper fall-off.

We also evaluated search performance relative to dimensions. As expected, there is a direct correlation between the two, which can be explained by the additional disk seeks to retrieve both cluster objects as well as semantic vectors for comparison to the query vector. This is illustrated in Figure 3 below.

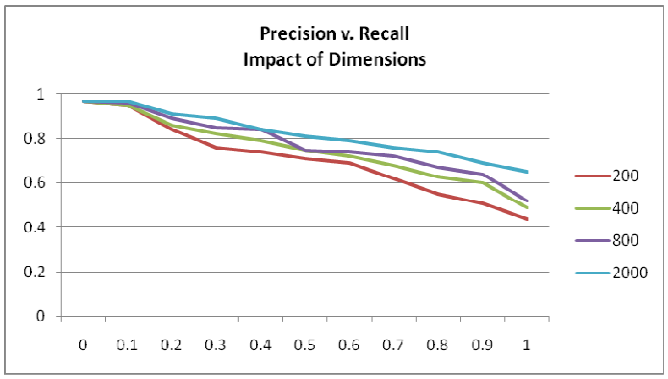


Figure 2: Precision and Recall graphs for the EDRM Enron Dataset

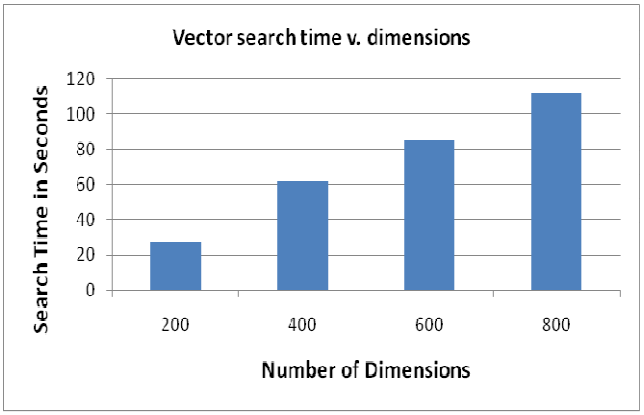


Figure 3: Characterizing Search time and dimensions for 20 random searches

A significant observation is that overall resource consumption increases substantially with increase in dimensions. Additionally, vector-based retrieval also times increase significantly. We need to consider these resource needs in the context of improvements in search recall and precision quality measures.

6.5 Discounted Cumulative Gain

In addition to Precision and Recall, we evaluated the Discounted Cumulative Gain (DCG), which is a measure of how effective the concept search related terms are [14]. It measures the relative usefulness of a concept search related term, based on its position in the result list. Given that Concept Search query produces a set of related terms and that a typical user would focus more on the higher-ranked entries, the relative position of related terms is a very significant metric.

Figure 4 illustrates the DCG measured for the EDRM Enron Dataset for a set of 20 representative searches, for four dimension choices indicated.

We evaluated the retrieval quality improvements in the context of increases in resource needs and conclude that acceptable quality is achievable even with a dimension of 200.

6.6 Impact of Training Cycles

We studied the impact of training cycles on our results. A training cycle captures the co-occurrence vectors computed in one cycle to feed into the next cycle as input vectors. As noted earlier, the document vectors for each training cycle start with randomly assigned signatures, and each successive training cycle utilizes the learned term semantic vectors and feeds it into the final document vectors for that phase. This new set of document vectors forms the input (instead of the random signatures) for the next iteration of the training cycle.

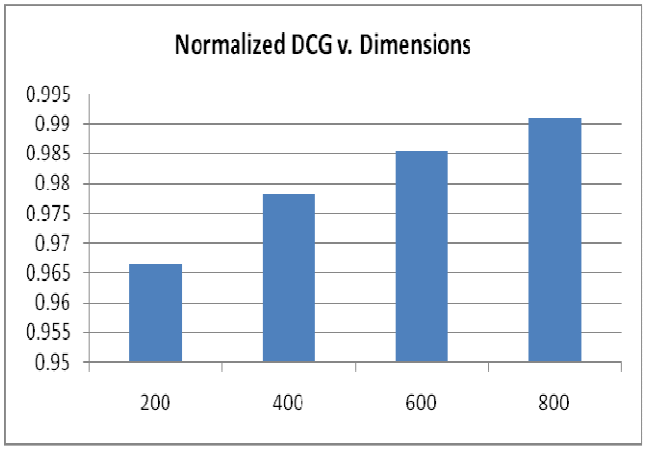


Figure 4: Normalized DCG vs. dimensions of semantic vector space

In our model, we note that term has a direct reference to another discovered term when they both appear in the same document. If they do not appear in the same document but are connected by one or more other common terms between the two documents, we categorize that as an indirect reference.

Adding training cycles has the effect of discovering new indirect references from one term to another term, while also boosting the impact of common co-occurrence. As an example, Table 6 illustrates training cycle 1 and training cycle 4 results for the term drilling. Notice that new terms appear whose co-occurrence is reinforced by several indirect references.

Another view into the relative changes to term-term similarity across training cycles is shown below. Table 7 illustrates the progression of term similarity as we increase the number of training cycles. Based on our observations, the term-term similarity settles into a reasonable range in just two cycles, and additional cycles do not offer any significant benefit.

Also noteworthy is that although the initial assignments are random, the discovered terms settle into a predictable collection of co-occurrence relationship, reinforcing the notion that initial random assignment of document vectors get subsumed by real corpus-based co-occurrence effects.

Query: drilling			
Training Cycle 1		**Training Cycle 4**	
Related Term	**Similarity**	**Related Term**	**Similarity**
Wells	0.164588	rigs	0.25300
Rigs	0.151399	wells	0.23867
viking	0.133421	offshore	0.22940
Rig	0.130347	rig	0.21610
buckeye	0.128801	exploration	0.21397
Drill	0.124669	geo	0.20181
exploration	0.123967	mcn	0.19312
richner	0.122284	producing	0.18966
producing	0.121846	ctg	0.18904
alpine	0.116825	gulf	0.17324

Table 6: Training Cycle Comparisons

Term1	Term2	TC-1	TC-2	TC-3	TC-4
offshore	drilling	0.2825	0.9453	0.9931	0.9981
governor	davis	0.3669	0.9395	0.9758	0.9905
brownout	power	0.0431	0.7255	0.9123	0.9648
brownout	ziemianek	0.5971	0.9715	0.9985	0.9995

Table 7: Term Similarity of training cycles (TC) for four cycles

7. CONCLUSIONS

Our empirical study of Reflective Random Indexing indicates that it is suitable for constructing a semantic space for analyzing large text corpora. Such a semantic space has the potential to augment traditional keyword-based searching with related terms as part of query expansion. Co-occurrence patterns of terms within documents are captured in a way that facilitates very easy query construction and usage. We also observed the presence of several direct and indirect co-occurrence associations, which is useful in a concept based retrieval of text documents in the context of electronic discovery. We studied the impact of dimensions and training cycles, and our validations indicate that a choice of 200 dimensions and two training cycles produced acceptable results.

8. REFERENCES

[1] Donna Harman, Towards Interactive Query Expansion, Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, Maryland

[2] Myaeng, S. H., & Li, M. (1992). Building Term Clusters by Acquiring Lexical Semantics from a Corpus. In Y. Yesha (Ed.), CIKM-92, (pp. 130-137). Baltimore, MD: ISMM.

[3] Susan Gauch and Meng Kam Chong, Automatic Word Similarity Detection for TREC 4 Query Expansion, Electrical Engineering and Computer Science, University of Kansas

[4] Yonggang Qiu, H.P.Frei, Concept-Based Query Expansion, Swiss Federal Institute of Technology, Zurich, Switzerland

[5] Ian Ruthven, Re-examining the Potential Effectiveness of Interactive Query Expansion, Department of Computer and Information Sciences, University of Strathclyde, Glasgow

[6] An Introduction to Random Indexing, Magnus Sahlgren, SICS, Swedish Institute of Computer Science.

[7] Trevor Cohen (Center for Cognitive Informatics and Decision Making, School of Health Information Sciences, University of Texas), Roger Schvaneveldt (Applied Psychology Unit, Arizona State University), Dominic Widdows (Google Inc., USA)

[8] Blair, D.C. & Moran M.E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM, 28, 298-299

[9] Berry, Michael W.; Browne (October 2005). "Email Surveillance Using Non-negative Matrix Factorization". Computational & Mathematical Organization Theory 11 (3): 249–264. doi:10.1007/s10588-005-5380-5.

[10] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman (1990). "Indexing by Latent Semantic Analysis" (PDF). Journal of the American Society for Information Science 41 (6): 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9. http://lsi.research.telcordia.com/lsi/papers/JASIS90.pdf. Original article where the model was first exposed.

[11] Widdows D, Ferraro K. Semantic vectors: a scalable open source package and online technology management application. In: 6th International conference on language resources and evaluation (LREC); 2008.

[12] EDRM Enron Dataset, http://edrm.net/resources/data-sets/enron-data-set-files

[13] Precision and Recall explained, http://en.wikipedia.org/wiki/Precision_and_recall

[14] Discounted Cumulative Gain, http://en.wikipedia.org/wiki/Discounted_cumulative_gain

[15] Latent Dirichlet Allocation, http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

RECOMMIND®
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).

*Manuscript submitted to the Organizing Committee of the **Fourth DESI Workshop on Setting Standards for Electronically Stored Information in Discovery Proceedings** on April 20, 2011. Updated May 18, 2011.*

Application of Simple Random Sampling[1] (SRS) in eDiscovery

Doug Stewart
Daegis

Abstract

eDiscovery thought leadership organizations advocate for the use of sampling throughout much of the EDRM process. Additionally, judging from the numerous and frequent references to "sampling" found in the eDiscovery literature and online content, there appears to be wide acceptance of the use of these techniques to validate eDiscovery efforts. At the same time, there are lingering questions and concerns about the appropriateness of applying random sampling techniques to eDiscovery data sets. This paper offers evidence that random sampling of eDiscovery data sets yields results consistent with well established statistical principles. It shows that Simple Random Sampling (SRS) can be used to accurately make predictions about the composition of eDiscovery data sets and thus validate eDiscovery processes.

Introduction

Sampling is often mentioned as the principal method of validating many eDiscovery activities and decisions. Thought leadership organizations such as The Sedona Conference, EDRM and TREC Legal Track have published guides, protocols and reports that explicitly call for the use of sampling techniques in various eDiscovery processes[2]. Also "sampling" is frequently mentioned in the literature, at conferences and in various forms of online content[3] as a key tool for validating results of collection, search, document review and other technology assisted eDiscovery activities. Further, the courts have called for the use of sampling in the eDiscovery process[4].

Despite these strong endorsements, there appears to be some reluctance or inertia toward the adoption and integration of sampling methods into the eDiscovery workflow. To some extent this reluctance may be based on a lack of understanding as most lawyers do not receive training in statistical principles. Lack of understanding may also contribute to the lingering doubts about the suitability of using Simple Random Sampling (SRS) techniques in the eDiscovery process. Additional education and training focused on applying sampling techniques in the eDiscovery process should drive adoption and acceptance of these methods. The Sedona Conference, EDRM and others[5] recognize this need and have provided leadership and advocacy in this area. Additionally, simple demonstrations that these techniques work may prove to be one of the best ways to dispel some of the concerns.

---

[1] A sampling technique where every document in the population has an equal chance of being selected.
[2] See http://www.thesedonaconference.org/content/miscFiles/Achieving_Quality.pdf; http://edrm.net/resources/guides/edrm-search-guide; and http://trec-legal.umiacs.umd.edu/LegalOverview09.pdf
[3] For example, "Using Predictive Coding – What's in the Black Box?" K. Schieneman et al. http://www.esibytes.com/?p=1649
[4] Victor Stanley, Inc. v. Creative Pipe, Inc., 2008 WL 2221841 (D. Md. May 29, 2008).
[5] "Sampling for Dummies: Applying Measurement Techniques in eDiscovery" Webinar by M. Grossman and G. Cormack 01/27/2011

This study sets out to test the efficacy and applicability of SRS techniques to the eDiscovery process. In doing so, it guides the reader through the process of applying sampling methods on eDiscovery data sets. Several sampling methods are described and tested. Additionally, the key parameters including sample size, confidence level and confidence interval are discussed and measured.

Methods and Material

The metadata of six inactive eDiscovery databases was searched and sampled for the purposes of this study. The databases ranged in size from a few thousand to more than a million records. Various fields including author, custodian, date, file type, and responsive were searched and sampled using the following four sampling techniques:

1. **Simple Random Sampling:** Random sample sets created by randomly selecting records from the specified population using the Microsoft .NET 3.5 Random Class to generate random record sets. Required sample size was one of the input parameters.
2. **Systematic Sampling:** Random sample sets created by selecting every n^{th} record from the specified population using a t-SQL script. A calculation was performed to determine the required value of n to produce the appropriate sample size.
3. **MD5 Hash Value Sampling:** Random sample sets created by running a MS SQL Server query to select all records with MD5 hash values beginning with two designated characters (e.g., AF or 4A). This method was used to produce a random sampling of $1/256^{th}$ of the population.
4. **Non-Random Sampling:** Non-random sample sets created by running a search for documents that fell within a certain date range. Not to be confused with a weighted sample.

The key parameters used to create the random samples for this study included:

1. **Confidence Interval:** Also called the "margin of error", the Confidence Interval indicates the precision of the sample's estimate by providing upper and lower limits on the estimate (e.g., plus or minus 2%).
2. **Confidence Level:** An indication of how certain one can be about the results. A 95% confidence level means that 95 times out of 100 the estimate will reflect the population's composition within the margin of error provided by the Confidence Interval.
3. **Sample Size:** Determined by using a sample size calculator. Required inputs include the desired Confidence Level and the desired Confidence Interval. The Sample Size is related to the Population Size but does not scale linearly. For example, the required Sample Size needed to achieve a 95% confidence level with a +/-2 % confidence interval is shown below for a variety of Population Sizes:

Population	Sample Size
1,000	706
10,000	1,936
100,000	2,345
1,000,000	2,395
10,000,000	2,400

4. **Population or Population Size:** The total number of documents in the source data set.

5. **Percentage or Prevalence:** The percentage of documents in the population that have the property being measured (e.g., percentage of the documents that are responsive). If the value is known it can be used to fine tune the Confidence Interval. If not known then 50% must be used to provide the most accurate estimates.

Sample sizes, confidence levels and confidence intervals were calculated using the sample size calculator found at:

http://www.surveysystem.com/sscalc.htm

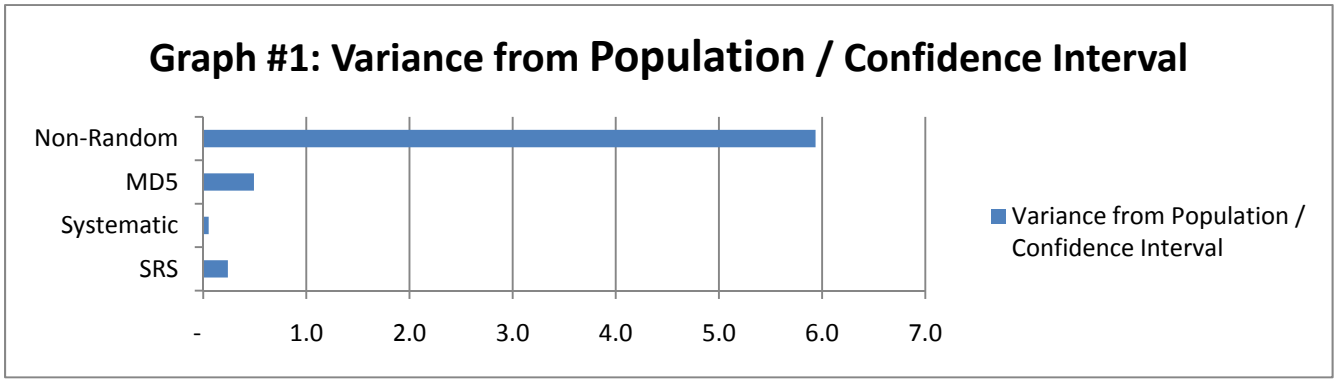
All analysis work was done using Microsoft Excel 2007.

Results

Graph #1: This graph shows the relative precision of each sampling method based on a single iteration of each. It shows how well the sampling techniques performed relative to each other. The precision is represented by the ratio of the absolute value of the sample's variance from the overall population for the property under investigation divided by the sample's confidence interval (or margin of error) as determined by using the sample size calculator. For instance, if the property under investigation were "ABC = Yes" the precision ratio would be calculated as follows:

Precision = abs((% of ABC = Yes in sample) – (% of ABC = Yes in population))/Sample's confidence interval

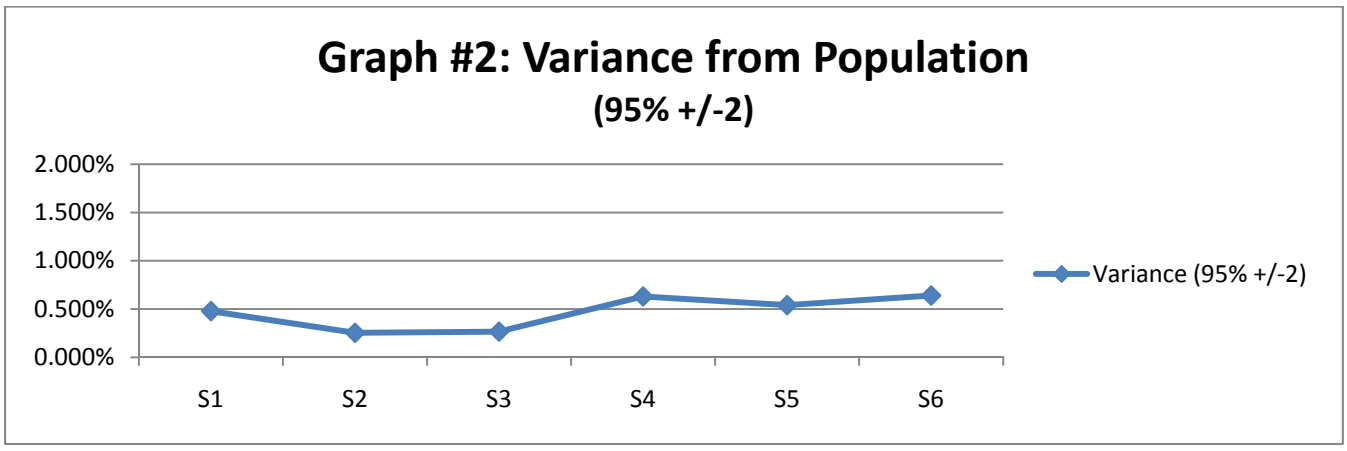
A result of 1 or less indicates the results fell within the confidence interval and thus indicates a sample that conforms to the principles of SRS and accurately characterizes the entire population. A result greater than 1 indicates a sample that does not accurately estimate the population. For example, precision score of 0.50 indicates the sample estimate varied from the actual population by half of the margin of error or confidence interval. A score of 5.0 indicates the sample estimate exceeded the margin of error by a factor of 5.



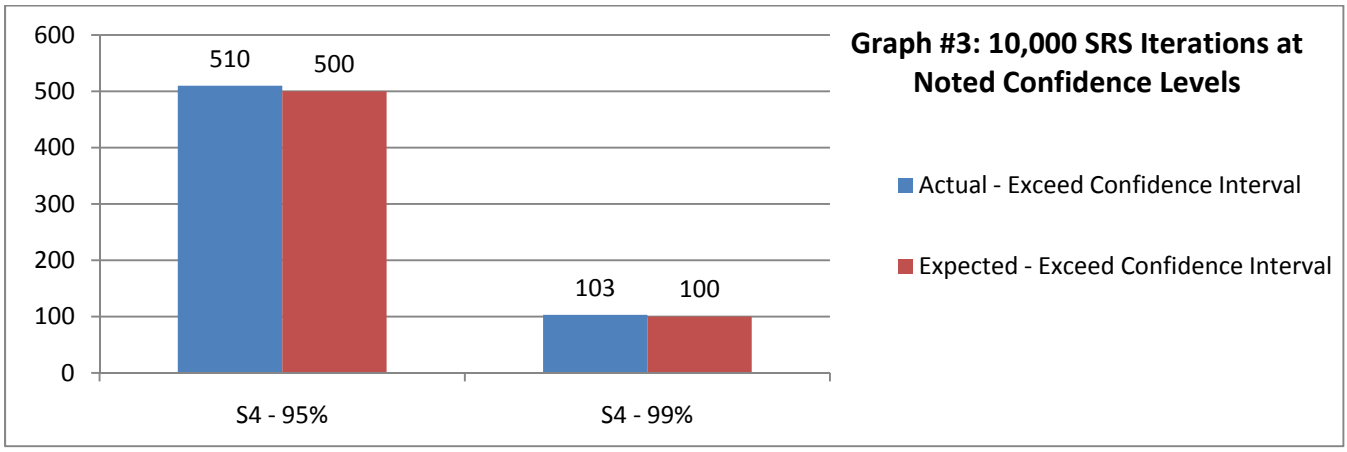
Graph #2: This graph shows the variance of the SRS derived sample from the population for six different eDiscovery databases. The sample size calculator was used to determine sample sizes based on a 95% confidence level and +/-2% confidence interval. The property analyzed was responsive (yes/no) that had been assigned in the review phase of each project's lifecycle. The variance was calculated as follows:

Variance = abs((% of Responsive = Yes in sample) – (% of Responsive = Yes in population))

The data sets (S1 to S6) ranged in size from approximately 4,000 to 1,400,000 records. The property under investigation ranged from an approximate 2% prevalence in the population to over 85% prevalence. The experimental data easily fit within the allowable margin of error.

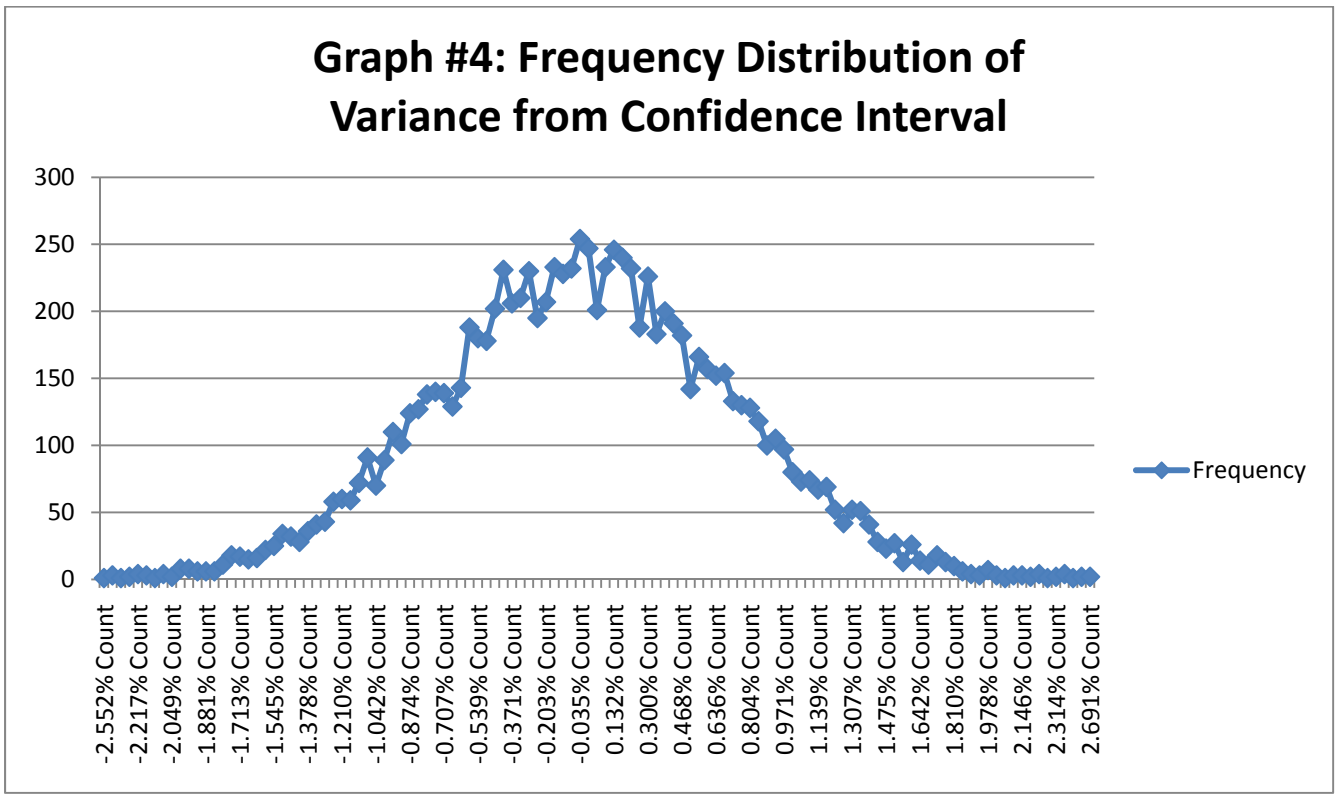


Graph #3: This graph shows the results of running 10,000 iterations of SRS on a single database two times and counting the number of samples that exceeded the confidence interval. The sample size calculator was used to calculate the confidence interval based on a specified sample size, confidence level and the known prevalence (percentage) of the record property under investigation. A confidence level of 95% predicts that 9,500 samples out of the 10,000 analyzed would produce an estimated prevalence that matched that of the population within the confidence interval range —500 (5%) samples would estimate a prevalence that fell outside the calculated confidence interval. A confidence level of 99% predicts that 9,900 samples out of the 10,000 analyzed would produce an estimated prevalence that matched that of the population within the confidence interval range—100 (1%) samples would estimate a prevalence that fell outside the calculated confidence interval. The experimental data match the SRS predictions with extraordinary accuracy.



Graph #4: This graph shows the results of running 10,000 iterations of SRS on a single database and then plotting the frequency distribution of each sample's percentage variance from the population. The sample size calculator was used to calculate the sample size based on the desired confidence level and confidence interval.

The data reveal that the distribution of the all the sample estimates centers on the actual prevalence percentage found in the population and then trails off as one moves out from the center as is predicted by SRS. As a result, this graph conforms to a normal distribution.



Discussion

The data represented in Graph #1 agree with established statistical principles and support the common assumption that random sampling techniques create samples that make more precise estimates or predictions about populations as a whole than non-random sampling techniques. In this study the non-random sample varied from the population by nearly six times the expected confidence interval or margin of error. The randomly generated samples all fell within the expected confidence interval.

Graph #2 demonstrates that SRS methods can be used across a variety of eDiscovery data sets to make predictions about the full population that fall within the calculated confidence intervals. The results shown indicate that regardless of the population size the SRS techniques were able to accurately estimate the population to within roughly 0.5 percent. The consistency in the accuracy of the estimates is even more astonishing when one considers that the prevalence of the property in question ranged from just over 2% to over 85% prevalence in the six data sets and the data sets themselves ranged in size from approximately 4,000 to 1,400,000 documents.

Graph #3 indicates that SRS of eDiscovery databases will produce results that fall within the calculated confidence levels and confidence intervals. The confidence levels are supported by the iteration data with remarkable accuracy—out of 10,000 iterations the results varied by only 10 samples and three samples from what was predicted by SRS.

The normal distribution seen in Graph #4 strongly suggests that SRS of eDiscovery data sets produces results that adhere to the well established statistical principles and body of knowledge. Specifically, the variance from the population for the 10,000 samples follows the distribution predicted by the Central Limit Theorem[6].

Conclusions

The prevailing assumption that SRS, when applied to eDiscovery data sets, produces results in line with accepted statistical principles is supported. This study provides compelling empirical evidence that supports the widely held belief that SRS is one of the best means of validating search and other eDiscovery activities.

The fact that a sample of fewer than 2,400 records from a population of one million can be used to accurately estimate the population as a whole may defy intuition. The best way to get comfortable with SRS is to employ the techniques and test them. Firsthand experience seems to be the best teacher.

Future work should include the creation of protocols and standards for further incorporating SRS methods into the eDiscovery workflow. This effort should also include standardized protocols for reporting on the sampling methods employed and the results obtained to ensure transparency in the process. Standardized protocols for the use of sampling techniques may also serve to educate and familiarize those that may have gaps in their understanding of these established techniques.

Sampling will play an increasingly important role in the eDiscovery process as the industry continues to mature, as data volumes continue to rise and as technology continues to advance. As such, the eDiscovery industry and thought leadership should continue their educational and training efforts to ensure that the relevant segment of the legal community is comfortable with the application of these techniques. Transparency in process, standardization, further training and practical demonstrations of how well sampling techniques work will go a long way toward achieving this goal.

Doug Stewart has over 25 years of IT, security and management expertise in the field of electronic discovery and litigation support. As Daegis' Director of Technology, Doug has been instrumental in the development and deployment of Daegis' eDiscovery Platform, which includes functionality for hosted review, on-site deployment, iterative search and much more. In 2009, Doug oversaw Daegis' ISO 27001 Certification for information security management, which includes a rigorous annual audit process. In addition, Doug manages several departments at Daegis including IT, data collection, and information security.

[6] The Central Limit Theorem states that as the sample size increases, the sample means tend to follow a normal distribution.