Automated Document Review in Litigation

By Sandra Serkes & Aaron Goodisman

Introduction

There are few things attorneys seem to enjoy less than a mind-numbing stretch of document review. For the uninitiated, document review is the process of systematically looking at each document of a large volume of case matter documents to determine whether they are important in some way to the case. Typically, the importance that attorneys are attempting to determine comes in three forms: privilege, responsiveness and topic area (also called issues or clustering).

When a document is privileged, it contains information that is legally protected from disclosure to other parties. Many times privilege rules are invoked to prevent the transmission of documents from one side to the other. This has the helpful effect of keeping certain information from being used at trial. However, because privilege is such a powerful tool in preventing access to information, it comes with a double-edge. If one party waives its privilege rights, either intentionally – or *unintentionally*- then the privileged information can now become fair game in the matter. (See Problems with the Current Approach, below.) In fact, accidental waiver of privilege is such an extremely damaging possibility, that many legal teams go to great lengths to prevent its occurrence.

Beyond privilege, comes the notion of responsiveness. When producing documents to another party, it is not enough to simply provide documents that are not privileged. The documents put forth must also be *relevant* to the matter at hand[1]. To determine relevance (also called responsiveness), the parties agree to limit the scope of the matter to certain topics, phrases and types of documents that most fit the bill of the issues in the case. For example, they may agree that marketing literature or emails about a company picnic are not relevant (thus, non-responsive). They may agree that patent filings or contractual agreements are highly responsive. Each case, of course, is different and much of the early case planning and preparations concern exactly what constitutes relevance to the matter.[2]

Finally, most legal teams find it helpful to organize and group documents in numerous ways. Some popular grouping techniques involve grouping by similar type of document, by similar document content or by similar topic area. Generally called “Issues” or “Clustering,” such groupings help the legal team

[1] Lest a sneaky party attempt to overwhelm its opponent with useless reams of documents with no bearing on the matter, something like a document-oriented filibuster.

[2] Late in 2006 the Federal Rules of Civil Procedure were amended specifically to account for such pre-discovery preparations and discussions. The rule changes make particular note of the need for cooperation and collaboration between opposing parties at this stage of the proceedings. See http://www.law.cornell.edu/rules/frcp/Rule26.htm for more information.

evaluate groups of documents at a time, rather than singly and randomly spaced throughout the population. Because clustering or grouping documents by similarity aids the determination of whether or not whole sets of documents meet the responsiveness criteria, smart legal teams use this helpful and cost-effective approach up front, in tandem with determination of privilege.

And so it has come to pass that any large-scale litigation matter invariably goes through a lengthy and costly process called Document Review. The greater the number of parties, complaints or years involved, the greater the scope of documents to review. With the proliferation of electronic means of communication (email, text messaging, blogging, tweets and so on), the number of documents in most matters is staggering[3]. While it is easy to remove the most egregious of the non-documents or the duplicative documents, substantial volumes of documents remain with the large majority ultimately undergoing some level of attorney review. The net result of which is that ungodly sums of money, time and personal effort are being spent on what will ultimately result in a handful of documents.

Problems with the Current Approach

The current approach of utilizing small armies of contract attorney labor to review documents on an outsourced or insourced basis is fraught with problems, some generally acknowledged and some hidden either due to the task's complexity, or the complicity of its purveyors.

Cost

To help combat the sheer volume of documents to be reviewed, many legal teams have turned to a form of temporary outsourcing, called contract (or outsourced) attorney review. The practice involves the use of temporary legal labor to assist in the document-by-document evaluations for privilege, responsiveness and issues. Contract review can take many forms. Some law firms or corporate legal departments add temporary staff on an hourly basis. Others look to outsourced, "managed" review companies that have cropped up across the US, in increasingly lower cost cities[4] (such as Detroit and St. Louis). Naturally, there are also offshore options in India, the Philippines and China. Each of these variations offers the option to provide legal analysis at lower and lower cost wage rates, with the current lowest option being on the order of ~ $10.00 per hour, as of Q2, 2011.

Disenfranchisement

As the document review work moves ever further away from the legal team with the case matter knowledge, the task must, by definition, become more routinized and less specific. Increasingly, the team performing the actual document review is located half-way around the world and is several steps removed from the client and its outside counsel. To help prevent miscommunication and mistakes

[3] For an excellent and visceral understanding of just how much document and data explosion there is, visit http://www.emc.com/digital_universe/downloads/web/personal-ticker.htm, a website from EMC that helps you calculate how much textual data you personally cause to exist each day. The writers of this white paper *each* create 2 GB every 24 hours!

[4] Interestingly, Tennessee, Georgia and Wyoming all have minimum wages *below* the federal rate, given a minimum set of conditions. See http://en.wikipedia.org/wiki/List_of_U.S._minimum_wages for state-by-state details.

(both of which ultimately cost more in time and expense), there has been increasing need for project management and systemization of the activity. This has led much of the outsourced document review to be separated into two tiers: first-pass and second-pass.

First pass review is simpler and more accepting of error on the part of reviewers. It is conducted by the lowest acceptable wage earners as quickly as possible. Second pass review is often (though not exclusively) conducted by higher priced outside counsel who have more knowledge about the case issues and the strategy for trial. The idea is to get as much of the routine document review out of the way as quickly and cheaply as possible, so as to put the bulk of the available resources (budget and time) where it is needed most. Put simply, the notion is that whatever might be incorrect in the first pass will ultimately be corrected in the second. While this is a smart approach, it leads many legal teams down a path of false confidence in the results. No one is particularly checking the accuracy and consistency of either pass. There is simply division of labor into rote and complex, with no real integration or assessment of the work quality. (See Accuracy & Consistency, below.)

Time

Time is an extremely precious commodity in legal matters. Anything that reduces time to case knowledge or increases trial prep time is a blessing, and anything that delays these things is a curse. Time in document review is measured in documents/hour, which is the rate at which outsourced attorneys accomplish their task. Thus, one attorney reviewing 60 docs/hour is twice as effective as another reviewing only 30 docs/hour. The actual length of the documents is, unfortunately, not germane to the value the attorney provides. Thus, the shorter the length of the documents, the more effective an attorney may seem to be, regardless of his actual quality of work.

Since time in document review is really a rate (docs/hour), the absolute time a task takes is relative. If the total document population to be reviewed is 100,000 documents, and reviewers perform the task at an average rate of 60 documents per hour, then the total (absolute) time for the task is 1667 person-hours. To reduce the completion time, simply add more people. If 1600 contract attorneys are used, then the task could be accomplished in under 2 hours! The art, then, becomes to assess how many contract attorneys should be used to produce the work in the least time possible, while still being manageable and resulting in high quality work.

Calculating Per Document Costs From Per Hour Model			
Total Documents to Be Reviewed		100,000	docs
Contract Attorney Review Rate		60	docs/hour
Total Workload - Initial Pass		1,667	person-hours
Contract Attorney Cost Per Hour	$	35	
Contract Attorney Costs	**$**	**58,333**	
Percent of Docs QC'ed		20%	
Number of Docs QC'ed		20,000	docs
QC Rate		120	docs/hour
Total Workload - QC		167	person-hours
QC Cost Per Hour	$	45	
QC Costs	**$**	**7,500**	
Project Manager Hours		160	
(assumes team of 10 reviewers, 8 hr shifts, 20 days of work)			
Project Manager Cost Per Hour	$	65	
Project Manager Costs	**$**	**10,400**	
TOTAL COSTS	**$**	**76,233**	
Cost Per Document	**$**	**0.76**	

Unfortunately, this is an inherent win-loss scenario. There will always be an incentive to add more contract staff to make the total task go faster. Yet, each additional person adds complexity and inconsistency into the mix. There is no hard and fast rule as to what the optimum balance is. Nor is there any way to actually measure the quality drop in adding each additional person. In fact, the only

thing that can be measured is the hours spent. Thus, the balance typically tilts towards speed, rather than accuracy[5]. Without comprehensive quality-control, this tilted balance can easily run amok, resulting in a shoddy, though quickly produced, review.

Set & Forget

While cases grow and change over time, and documents are often being added into the mix on an ongoing basis, the current approach to document review treats the activity as a one-off, "just get through it" basis. Reviewers almost never go back and re-visit documents to change an assessment or re-think the strategy behind the document disposition. Yet, legal teams are constantly re-evaluating their position and approach as new facts come forward. Document Review strategy should match the evolving themes and priorities in the case and trial strategy. Ideally, it should be easy to re-review a document population with new issues and ideas, without losing much time or spending additional funds. Imagine if document review could be conducted the way one looks a light through a prism – different from every angle, yet easy and low-cost to reproduce and document.

Work Quality: Accuracy & Consistency

Perhaps the gravest issue currently plaguing large-scale document review is the work quality of the resulting output. For many years, the industry assumption has been that if an attorney, a person, looks at the document then he or she is capable of making sound judgements about the attributes of the document. While this is generally true on an individual document-by-document basis, the notion pretty much falls apart on an aggregate basis. People simply do not have the capacity for the near infinite complexity that a large document population holds. No one person can possibly repeat their analysis and decision-making for hundreds of thousands of documents, particularly when the decisions are taking place across a long period of time (typically weeks or months). Adding multiple people into this process only compounds the inconsistencies and errors.

Such problems have been present in document review from the beginning, but the smaller document populations and concentration of review activity near the case knowledge source helped reduce and mask it. There have been numerous studies that have documented beyond any doubt as to the inadequacies of linear review conducted by teams of people over time[6]. In one particular study, associates and paralegals assumed they had found 70% of the responsive documents, when in closer auditing and analysis, they had only found 20% in actuality[7]! For those in the business of documenting accuracy and consistency, we see this phenomenon over and over again.

---

[5] There is another more insidious incentive to move a project ever-faster. Everyone in the service provision chain can bill and ultimately be paid faster if the total project duration is shorter. For those operating on commissions or bonus structures, this can be a material incentive.

[6] Several well-known studies of this topic include those by the Text Retrieval Conference, Legal Track (TREC), the Electronic Discovery Institute and Blair & Maron. Valora Technologies has also conducted numerous performance evaluations of linear, human review vs. machine, software review. Results available upon request.

[7] The Blair & Maron study involved a manual review of 350,000 pages (40,000 documents), for the purpose of finding responsive documents with particular content. As noted, the lawyers in the study greatly overestimated their effectiveness at finding relevant documents.

Given the responsibility of outside counsel to stand by and certify the work product produced by any and all contract attorney labor[8], *any* inconsistent or inaccurate document review is a potentially significant liability to both the case and the attorney permitting it. The second half of this paper goes into extreme detail about how to measure and assess accuracy and consistency in document review, whether performed by people, by software, or some combination.

Confidentiality

Thankfully, it is standard practice to have outside contract attorneys sign non-disclosure agreements about the material they are privy to during the course of their assignments. However, many legal teams are overlooking 2 key potential problems: 1) enforceability and 2) differing privacy & disclosure laws in different countries. While we all hope that those in the legal profession hold themselves to a high work ethic and obedience of non-disclosure agreements, it is naïve to think that such practices are in perfect adherence across the board. It only takes one disgruntled temporary laborer to disclose a critical piece of information for the whole practice to be called into question, and rightfully so. In general, the fewer the number of people with access to privileged information, the better. This incentive is in direct conflict with the incentive to finish the task faster (see Time, above), creating a tension that has no positive outcome. As before, increased confidentiality is a "soft," hard to measure benefit, while faster time is quite concrete and easy to measure. Guess which one usually wins.

We question how many legal teams currently enjoying the fruits of low-cost, offshore labor rates for contract attorneys have actually done their homework on the privacy and non-disclosure laws of the country to which they are sending important client documents. It is well known that India, in particular, has loose, forgiving laws around US patent protections, generally favoring the party US laws would consider to be infringing. How many US-based companies are sending IP litigation documents to India, or Pakistan, or the Philippines? More importantly, as the reader of this paper, do YOU know where your or your clients' documents are going and are YOU familiar with the privacy and non-disclosure laws of that country?

Work Conditions

Since we cannot possibly know the working conditions of each and every document review project, we will refrain from broad blanket statements and keep this section short. Surely, some review projects are conducted in gleaming legal offices with all the amenities, but many are not. With low cost being the primary driver, there is a strong incentive for legal reviews to be conducted in low-rent, less-than-optimal office environments, with questionable working conditions for reviewers. We advise anyone participating in outsourced contract review activities to visit the websites[9] written by and for contract attorneys in the field. As to working conditions in other countries, we cannot even begin to speculate.

[8] See Federal Rules of Civil Procedure, Rule 26 (g) (1) (A) & Federal Rules of Evidence, Rule 502

[9] temporaryattorney.blogspot.com, www.theposselist.com & www.toiletlaw.com are several notable sites. Please note that the language and content of these sites can be very strong at times and Valora Technologies has no affiliation with or responsibility for the content found there. We mention them only in the spirit of letting the buyer beware.

The net result of the interplay of the problems listed above is that the current status quo seems to be a mantra from legal teams to outsourced attorneys: get this work done as soon as possible, with little regard for accuracy, consistency or working conditions and don't look back.

A better way..

Fortunately, a new approach to document review is emerging, fueled in part by advances in technology and in part by a growing desire of the judiciary to bring the proportionality of the expense of discovery back in line with the stakes of the litigation[10]. The rest of this paper discusses the merits and emergence of Automated Review, called AutoReview as created and implemented by Valora Technologies, Inc., an industry pioneer in the analysis and disposition of documents.

Defining AutoReview

AutoReview is the process of using computer software to determine the responsiveness and privilege designation of each document in a collection, whether paper or ESI in origin. AutoReview also includes the identification and tagging of documents for confidentiality level, grouping documents together based on attributes, marking documents with case-specific characteristics (issues), automated redactions and more.

There are several approaches to AutoReview. In this paper, we discuss the Rules-Based approach in use at Valora Technologies. By "Rules-Based," we mean that the software uses a set of rules to determine the characteristics of each document. The software applies these rules to the information available to it about each document, including, but not limited to, the content of the document (the document text), the metadata of the document (if available or derivable) and the context of the document (such as whether or not it is an attachment, a duplicate/near duplicate or its custodial origins). Each rule indicates that if a document has a certain set of features, then the software should flag it with a particular characteristic.

The set of rules used for AutoReview of a particular collection of documents within the context of a particular case are specific to that document collection and that case. Some of the rules may be common across many cases (such as Attorney-Client Privilege) or cases of a particular type (e.g., bankruptcy litigation). Rules may be common across many document collections or specific to a particular document collection. Part of Valora's AutoReview methodology includes the creation of an appropriate set of rules for each AutoReview project. An important characteristic of Valora's technology is that it makes the creation of this RuleSet straightforward and efficient. RuleSets are typically created within 72 hours or less.

[10] See "Asserting and Challenging Privilege Claims in Modern Litigation: The Facciola-Redgrave Framework," written by Judge John M. Facciola and Attorney Jonathan M. Redgrave for The Federal Courts Law Review in 2009.

Content Analysis Rules

As a simple example of a content analysis rule, consider determining whether or not a document is a letter, as opposed to some other type of correspondence (e.g., a fax) or a non-correspondence document (e.g., profit and loss statement). Although this rule might not be applicable to a review for responsiveness or privilege, it serves to illustrate the general approach of creating a rule that makes a determination about a document based on that document's content.

What distinguishes a letter from another type of document? Or, put another way, what clues allow a human being to determine that a document is a letter? A letter typically includes on the first page the sender's letterhead, a date, a block of text designating a recipient, and a salutatory phrase, such as "Dear Bob," or "To whom it may concern." This is followed by the body of the letter, which may extend for several pages, and then a closing phrase ("Yours truly") and information about the letter's author (name, title, etc.). This description of a letter can serve as a template for identifying letters within a collection of documents that contains both letters and non-letters.

Analysis Using Software

The preceding description of a letter can be represented so that a software system can analyze the text of a document to make a similar determination of whether or not it is a letter. For example, although the date of a document may appear in any of several formats (e.g., March 15th, 2011 or 3/15/11), it is straightforward and finite to represent the most common patterns of characters representing a date in a way that a software system can detect it.

A similar approach allows software to detect the name, title, company, address and phone number that make up the block of text designating the recipient. Just as the date recognition can allow for variations in the format of the date, the recipient block recognition can allow for variations as well. For example, the recipient block may or may not contain the recipient's title or phone number.

Similarly, a software system can detect the remainder of the elements that identify a letter using similar techniques of pattern recognition and textual analysis, and allowing for common variations. This collection of patterns forms a "rule" for identifying letter documents, as well as the letter's date, author(s), recipient(s), any copyees or blind copyees, and its title or subject line.

Recognition, Misrecognition and Iteration

Of course, there may be documents within a collection that technically match this rule, but are, in fact, not actually letters. If this occurs, a modification to the rule can make it more precise, to match fewer documents, more of which are letters.

Similarly, there may be documents within a collection that are letters, but that do not match this rule. Additional rules or an expansion of this rule can identify these letters. Through several rule refinements and additions, the performance of letter identification rules can become as accurate as is necessary.

Evaluating Rules: Sampling, Recall and Precision

A collection of rules (or "rule set") constitutes a recognition algorithm and as such can be evaluated on the extent to which it correctly identifies the document characteristics it targets.

In order to evaluate the accuracy of a rule set, it is necessary to have something to compare it to. That is, the software, using the rule set, generates for each document a set of characteristics. In the example above, the letter recognition rule set generates a determination for each document that it either is, or is not a letter. In order to evaluate the performance of the rule set, it is necessary to know the "truth"; that is, whether each document is actually a letter or not.

In creating and testing this system, Valora uses a collection of documents for which the "truth" is known in advance. We can then evaluate the system by comparing the system's results to this truth. In an actual case, no such truth exists. If we knew the correct set of characteristics for each document, there would be no point in running the system, *except* for evaluative purposes.

In developing a rule set for an actual case, therefore, we must create a set of truth values for a subset of the document collection and evaluate the rule set against that subset. This practice is called "sampling," because we are looking at only a sample of the document collection. Sampling is a statistical approach to learning something about the whole collection without actually looking at every member of the collection. There is a significant body of work that indicates that, with the right parameters, information derived from a statistical sample (subset) of a collection is with high probability the same as that of the whole collection.[11]

In this case, the information we are deriving is the accuracy of the rule set. To do this, we select an appropriate subset of the collection, determine the truth for each member of the subset, and then compare the results of running the rule set to that truth.

The selection of the subset is an important part of ensuring that the information we derive is representative of the whole collection. There are many approaches to creating a statistically valid subset, but all have two important characteristics: selection method and size. It is not sufficient to select, for example, the first documents in a collection. That set of documents might not be representative of the whole collection, and so the evaluation of the rule set might also not be representative. It could either be much higher (the selected documents happen to fit the rule set in such a way that it appears to be more accurate than it really is) or much lower (the selected documents all happen to be precisely the cases the rule doesn't handle, such that the rule appears to be less accurate than it really is). Although there are many methods of selecting subsets, the most common, reliable and easy to implement is random selection. By selecting documents randomly from within the collection, we maximize the likelihood that the subset of documents accurately represents the variation of documents within the whole collection[12].

[11] For more information on statistical sampling, the WikiPedia entry (http://en.wikipedia.org/wiki/Sampling_%28statistics%29) is a good place to start.

[12] An alternative approach that Valora sometimes uses is to create a representative (stratified) sample of the population. This "mini-me" sample is a microcosm of the total population, reflecting the same proportions of document and attributes as the whole. When available, representative sampling is the best option. To create a representative sample, Valora utilizes our companion technology, AutoCoding, to populate objective field data for many fields, such as Document Type, Authors & Recipients and document Date. By utilizing AutoCoding first, we can use its results to inform the stratified sample selection.

It is also important to select a subset of an appropriate size. Although it is true that a larger the subset provides a more accurate the evaluation of the rule set, there is a point after which the addition of more documents to the subset does not measurably improve the evaluation (but does increase the cost). For example, if we have considered only a single document, then looking at a second document would clearly increase our understanding of the accuracy of the rule set. If we have considered 100,000 documents, however, then the addition of one more document is unlikely to tell us anything we didn't already know.

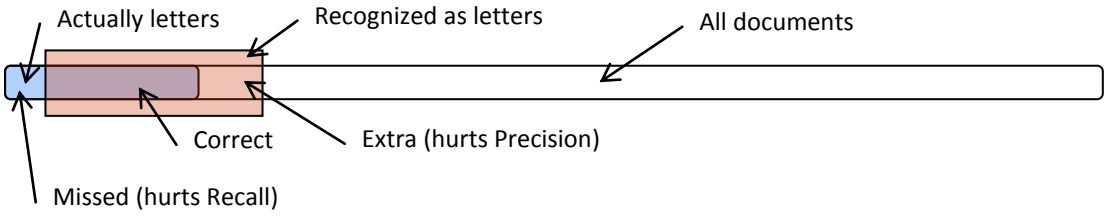
By making worst-case assumptions about the distribution of results within the collection and considering the size of the whole collection, we can compute sample sizes that provide whatever level of reliability to the evaluation we desire. At Valora, we choose sample sizes that give us 95% certainty that the evaluation of a rule set against the sample is within approximately 2% of what we would find were we to evaluate the rule set against the entire population.[13]

Comparing Results to a Sample

A naïve approach to evaluating the accuracy of a rule set would be to compare the rule set's results to the truth and compute the percentage of documents for which the rule set produced the correct answer. The problem with this approach is that it does a poor job of evaluating sparse data[14]; that is, the situation in which only a few documents are flagged with a particular characteristic.

Suppose, for example, that in a sample of 100 documents, only five are actually letters. Now suppose the letter recognition rule set fails to identify any of the letters. That is, the rule set indicates that none of the 100 documents are letters. How accurate is it? If we simply compare the rule set's results against the truth, we find that the rule set is correct 95% of the time. For 95 of the documents, the rule set says they are not letters and it is correct. But, despite that high score, the rule set has done nothing useful, since it has failed to find the letters.

Instead of this document-by-document scoring, we use a pair of statistical measures commonly selected for evaluation of recognition algorithms called Recall and Precision. These measures are specifically designed for evaluation of sparse data sets because they ask two important questions: How many of the documents the rule set was supposed to find did it actually find? (That's Recall.) And how much of what the rule set found was it supposed to find? (That's Precision.)



Example of Precision and Recall

[13] To be more precise, the statistical sample size predicts that the result we get is within our target percentage of what we would find were we to evaluate the rule set against *any other randomly selected subset*.

[14] Sparse data (< 20% of the documents have the attribute) is a frequent occurrence in litigation. Typically, fewer than 20% of documents collected end up being responsive or privileged and < 10% have any particular issue code.

Consider these separately. In the example above, there are five letters in the collection of 100 documents. If the rule set identified four of those five documents as letters, then it has a Recall of four out of five, or 80%. That is, the rule set found 80% of the letters it was supposed to. Of course, it is easy to construct a rule set that has perfect Recall. If the rule set marks all 100 documents as letters, then it certainly marked the five actual letters correctly, so its Recall is five out of five, or 100%. But the rule has done nothing useful, since it has not helped us identify the letters within the collection.

Precision measures how much of what the rule set found was correct. Suppose the rule set actually marked six documents as letters, but only four of them were actual letters. Its Precision score would be four out of six documents, or 66%.

We can average these two measures, Recall and Precision, to create an overall evaluation of a rule set[15], but in performing review, it is better to keep these separate, because they have different impacts on litigation.

Recall vs. Precision in AutoReview

In review, Responsiveness is the characteristic of a document that indicates that it is relevant to a request for document production and should thus be produced to the opposing counsel. There are legal consequences for inadvertently not producing all of the required documents as well as for producing more than the required documents. To use the terms of statistical measure, there are consequences for poor Recall (not producing all the required documents) and for poor Precision (producing extra documents).

These consequences are not equal, however. In litigation, there are generally only legal ramifications for overproduction (poor Precision) when the overproduction is egregious and with the intent to obfuscate the material; that is, to bury the opposing counsel in irrelevant documents. Conversely, the ramifications are more serious for withholding documents that should have been produced (poor Recall). Thus, for Responsiveness, Recall is a more important measure of rule set accuracy. That is, it is more important that a rule set have high Recall of Responsiveness (find all the responsive documents) than that it have a high Precision (find *only* responsive documents).

Similarly, the ramifications for Recall and Precision for Privilege designation also favor Recall over Precision. Inadvertent production of a privileged document (poor Recall) can have serious consequences in waiving the right to privilege for that document and others. Inadvertent withholding of a document, particularly in a clearly defined, defensible manner, is a more easily remedied situation.

The same dynamics hold for issue identification. The purpose of characterizing documents as being relevant to particular case issues is to help the case team use the material within those documents to formulate and execute their case strategy. An issue characteristic does this by identifying a subset of the documents that relate to a particular aspect of the case, and allows the case team to go through that subset in more detail. Missing documents related to an issue (poor Recall) is a more serious

[15] These are commonly combined using a harmonic mean (a kind of average) called the F-score or F-measure.

problem for the case team, because they will not have the benefit of using those documents for their analysis, than incorrectly identifying extra documents as related to the issue (poor Precision).

Rules for Responsiveness, Privilege and Issues

The rules-based approach to document content analysis described earlier for identifying letters can also determine document responsiveness, privilege and issue designation for a particular discovery project. This is slightly different from the letter example in that what constitutes a letter does not change significantly from one document collection to another, but what constitutes Responsiveness depends on the request for documents as well as on the documents themselves. A document might easily be responsive to one request, but not to another.

Although the rules for Responsiveness, Privilege and Issues do not transfer completely from one discovery project to another, the task of using software to determine them is the same as the process used to identify letters. The task of identifying these document characteristics is broken down into subtasks of identifying those document features that indicate each characteristic and then representing these to be used by a software system.

Common Patterns for Responsiveness, Privilege and Issue Rules

Rules can be arbitrarily complex for identifying documents that are Responsive to a particular request, Privileged within the context of a particular litigation or that correspond to particular issues within a case. Nonetheless, these rules frequently follow certain patterns.

In general, documents are requested according to author, date range and topical components. That is, Responsiveness typically encompasses documents that fall within some period of time, that are authored by, received by, or communicated to a particular group of people, and that discuss certain case-specific topics. Responsiveness might also be determined in part by the type of document. For example, financial documents might be either definitively Responsive or non-responsive as a group for a particular case. Accordingly, a good starting point for creating a rule set for Responsiveness is to identify the who, what and when of the documents being requested, as best as this can be determined at the outset of the review[16].

Similarly, documents are withheld for privilege typically because they are attorney/client communication, because they are attorney work product regarding the case, or because they contain trade secret or other protected material. Attorney client communication is by definition between a limited set of people, so author and recipient information is key in determining this. Attorney work product is typically authored by one of the attorneys relevant to the case. And trade secret material typically contains key phrases or falls within specific document types (e.g., patent application). These components combine together to form the basis for the rules that designate documents as Privileged.

Marking documents as being relevant to particular issues within a case is also typically a function of

[16] Valora typically utilizes metadata when available, or AutoCoding for this purpose. (For more on AutoCoding, download our white papers available on www.valoratech.com.)

- particular key phrases
- types of data (e.g., patient numbers, manufacturing component designations) or
- types of documents (e.g., sales and marketing brochures).

Valora's process for creating AutoReview rule sets typically starts with these frameworks.

Valora's Process for AutoReview

Valora's process for AutoReview begins by receiving a collection of documents along with the parameters of the review project. The documents are received as text, images and/or meta-data[17] if available, and any load files used to designate document boundaries and to connect images and text with each other and with meta-data and native files. The parameters of the review project include any written review guidance the case team may have prepared (often a review memo or protocol or a coding manual) as well as the original document requests received from opposing counsel.

Valora's Technical Operations staff load the documents into Valora's PowerHouse document analytics system and verify the integrity of the document collection, ensuring that all documents listed are present, that each document has text and images, that meta-data is formatted as expected, etc. Documents that have images, but are missing text are run through OCR software.[18] Documents that are missing meta-data may be run through Valora's AutoCoding service, which assigns information such as document type, date, title, author, recipient, etc. based on generic and collection-specific document recognition rules.

While the document collection is being loaded and evaluated, Valora's project management team goes over the review guidance with a representative of the case team. This helps flesh out the review guidance with any ideas the case team may already have about what document features correspond to the review designations and allows the case team to go over any example documents they may already have reviewed.

From this point the Valora review team, including the Project Manager and Technical Operations Specialist, begin to create a human-readable version of the rule set for each review designation. They begin by performing basic searches on the document text and meta-data fields for key terms and likely phrases based on the review guidance and discussions with the case team. Looking at the documents that match these searches, they identify:

- Indicative, confidence-boosting document features; that is, features of the documents other than the search terms that serve to confirm the hypothesis that a particular document meets the target criterion
- Contraindicating document features; that is, for documents that match a search, but don't match the target review designation, features that can indicate that situation, and

[17] Meta-data refers to any electronically extracted information about a document. For example, most processing of litigation documents extracts authors, recipients, subject, etc. from any email documents into a fielded data file.
[18] Optical Character Recognition software converts pictures of textual documents into computer-readable text

- Other features that the matching documents have that could be used for further searches; that is, features that the matching documents have in addition to the search terms that other target documents might have without the search terms.

Example of Indicators & ContraIndicators

- *Indicator: Google, Inc. as a party to litigation (perhaps as Author of a patent)*
- *ContraIndicator: Google as a gmail.com address on email.*

Once the team has assembled an initial collection of rule tests, the Technical Operations specialist represents these patterns in Valora's rule definition system and applies this rule set to the collection of documents. This whole process of creating the initial rule set criteria, encoding it for software use and evaluating it on all the documents typically takes place all during the first day of analysis and processing.

Once the processing is complete, the Valora team evaluates the results along two dimensions: First, is the rule set correctly identifying the documents the team intended it to? That is, is the rule set implemented correctly? Second, and more importantly, which document/characteristic combinations is the rule set incorrectly identifying (with either poor Recall or poor Precision)? The team creates statistical samplings of the document collection and evaluates each document and characteristic to identify both incorrect and missed values. More importantly than simply evaluating the rule set results, however, the team uses this process to identify the cause of each type of error and to adjust the rule set accordingly, either to include additional documents for a particular characteristic or exclude documents, based on additional features identified within the documents with incorrect results.

In addition, this process often generates questions that the Project Manager takes back to the case team for clarification. This is also an opportunity for the project manager to provide the case team with valuable information about the makeup of the document collection.

As part of this discussion between the Project Manager and case team, this interim state of the collection and review results can be loaded into Valora's BlackCat data visualization system. This system, designed to work in concert with the PowerHouse analytics engine, provides a web-based view on the document collection and the results of the analysis.[19] The case team can easily see how different characteristics and issues are distributed across different subsets of the population; for example, how many of the documents from 2008 are Responsive, or in how many documents two issues both occur.

The Technical Operations Specialist then updates the rule set based on this feedback and runs the collection again. The Valora team repeats this process iteratively until the results have sufficient accuracy.

As part of this process, the document collection is divided into three subsets:

- Documents that are confidently characterized by the system

[19] For more information on the BlackCat system, please see Valora's web site at www.valoratech.com or contact our sales team.

- Non-automatable documents that cannot be characterized by the system at all (because they have no text or are otherwise not computer-readable)
- Documents that appear to be ambiguous; that is, they match both confirming and contraindicating features of one or more rules.

Documents in the last two categories are flagged for potential additional processing as error conditions. Depending on the case team's strategy, budget, etc., these documents are reviewed in a more traditional manner by Valora staff or by the case team themselves. Typically, less than 20% of the total document population requires such additional processing.

When the iterative process is complete, the automated results have achieved the desired accuracy and the uncharacterized documents are flagged (and potentially evaluated manually), the data is exported from the system in any of several standard formats and delivered per the case team's instructions. Many Valora clients choose to receive the completed data as a load file for their review platform of choice (often Relativity, Ringtail, iConect, Concordance or Summation).

Underlying Technology

Valora's rule-based content analytics system (PowerHouse) is based on a hierarchy of text recognition algorithms. As described above in a top-down fashion for the identification of letters, each rule consists of specified patterns of items, each of which are themselves simpler patterns. For example, the letter rule indicates a date followed by a recipient block. The date portion of the pattern is itself a pattern, consisting of a month, a day and year. Each pattern can contain various kinds and levels of flexibility, allowing for the presence or absence of sub-patterns, variations of sub-patterns (multiple options) and different orderings of sub-patterns. That is, a letter rule could allow the date to be missing or to follow the recipient block instead of preceding it. The date pattern could allow the month to be a number, word or abbreviation.

The PowerHouse system provides many built-in patterns (such as dates, people's names, addresses, etc.) as well as pattern combination operators, such as boolean operations (and, or, not) and proximity (within a certain distance in the text). In addition, the system provides a description language and interface that allows the Technical Operations Specialist to represent complex rules quickly and easily and to make those rules visible and easy to expand or modify.

The system distributes the application of rules to documents across multiple virtual computer systems, allowing large collections to be processed in parallel. This allows the Valora review team to get results for a collection quickly and to iterate the process of rule evaluation and refinement repeatedly within a short timeframe. The Technical Operations Specialist monitors the processing from a central console application that shows the disposition of the collection within the system as well as the state of each computer system throughout the process.

Redaction

Frequently, an entire document is either Responsive to a request or it is not; the document is Privileged or it is not. Sometimes, however, only a portion of a document is appropriate to produce, such as with a portion of a long string of emails or in a report covering status of numerous drug under development. In this case, the rest of the document is removed (blacked out) and the document is produced in its protected state. The process of removing portions of a document prior to production is called Redaction. Portions of a document may be redacted because they are privileged, because they contain privately identifying information (PII) or because they are protected by trade secret or other rules.

The technologies described above used to flag whole documents in litigation review can also identify and redact portions of documents. Valora's AutoRedaction services use the same types of pattern recognition technologies to identify and mark sections of documents to be redacted. These sections are then either automatically blacked out, or delivered to the case team as proposed redactions for further review.

Additional discussion of AutoRedaction is beyond the scope of this paper, but for more information, please contact Valora directly.

Defensibility

The defensibility to the court of this configurable, rules-based approach to document review rests on the facts that:

- it is consistent
- it can be explained
- it can be evaluated and
- it can be repeated.

Because this method of review analyzes documents based on their content, documents with identical content are guaranteed to be reviewed identically, and similar documents similarly.

The case team can describe the rules used to determine document Responsiveness and Privilege in as much detail as is necessary, including precisely which authors, recipients or key terms were used in determining each characteristic. If appropriate, the system can generate a complete, hierarchical rendering of the rule set, although this usually provides much more detail than is appropriate for a court setting.

The case team can, if necessary, produce statistical samples with Recall and Precision scores from the rule set evaluation that is part of the standard process to demonstrate that the rule set accurately captures the desired document characteristics. They can also perform separate, independent quality assurance measures.

Most importantly, because processing the collection using the rules takes only a short time, the case team can point to the fact that that performing the review again would produce the identical results, a

claim that cannot always be made about other approaches to review, notably the manual, document-by-document review.

Lastly, it is worth noting that comparisons between manual reviews and automated review techniques, including this one, have demonstrated that the automated review results are superior[20].

Conclusion

The time has come that well-managed and configurable software systems can outperform large groups of contract attorneys for first pass assessments of privilege, responsiveness and issues – what is typically known as Document Review. With defensible processes and accurate software configurations, such results can be easily created, repeated and documented for far less money, time and effort than a traditional, manual, linear review. In the spirit of efficiency, ethics and proportionality, we urge legal teams with large document populations to consider the benefits and cost savings of utilizing an automated approach to document review.

[20] Interested parties should review the results of the Text Retrieval Conference, Legal Track, which focuses on rigorous, academic assessment of automated means of performing document review against manual means. See http://trec-legal.umiacs.umd.edu/LessonsLearned.pdf to start.