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