

Data Analytics: e-Discovery Accuracy, Defensibility and Cost Efficiency

By Eugene Eames

The explosion of electronic documents in the corporate environment has had a profound effect on the litigation-discovery process. With computers pervasive in our society, the number of electronic documents that must be examined for responsiveness is usually enormous.

Because it is almost never cost effective — or even possible — for attorneys and paralegals to review every document, litigation-service providers use a variety of electronic tools to reduce the number of documents to review.

But one should keep in mind that keyword-based automated search tools are not necessarily accurate, especially when the search terms are brainstormed by counsel in a vacuum. And new-age concept-based tools, while often quite effective at targeting documents based on subject matter or concept, are technologically hard to explain and, as a result, hard to defend.

Data analytics adds testing and analysis to keyword-based search, providing litigators with more accuracy and defensibility — often at a much lower cost.

SOME NUTS AND BOLTS

Every discovery project starts with the collection of huge volumes of electronic

data, and results in attorneys or paralegals determining which documents are responsive and must be produced to opponents. The in-between steps that electronic-discovery providers take usually involve preparing and organizing the documents for review, including filtering certain file types (such as applications) and duplicates. For years now, many e-discovery projects have also included a keyword-search component to help reduce the document collection to a manageable size for review.

Typically, the attorneys brainstorm keywords based on general terms that they believe will be present in responsive documents and the search mechanisms extract documents that meet these selection criteria. While Boolean search techniques (such as selecting documents that have only keywords within a certain proximity to other keywords) make this process somewhat accurate, it should include a procedure for testing and refining the search strategy to increase accuracy, resulting in a more defensible process and a lower overall cost.

SEARCH TERMS EVOLVED

In a data analytics-based process, counsel still typically brainstorm initial search terms, but those terms are then tested and researched within samples of the collected data, researched in repositories of information available to the public on the Web and considered in the context of any litigation pleadings. This method allows the accumulation of subject-matter knowledge and provides the legal team additional insights as it refines its search terms.

Once the legal team agrees on initial search terms, those terms are run against a small percentage of collected documents. The returned documents are then reviewed by the legal and data-analytics teams to determine the effectiveness of the search criteria. Advanced concept and text-analysis tools also might be applied to samples of the collected data to help determine possible alternative search strategies, such as introducing potential modifications to the search terms that can then be tested and validated. This process might be repeated many times until the teams are quite confident about the selection criteria. Ultimately, all search criteria are depicted in standard text-search or relational-database search syntax. By using standardized syntax, the selection criteria become easier to explain, even though they might have been developed using the findings of advanced tools and techniques.

For example, in a recent case involving a large pharmaceutical company, we started by testing 5% of data samples from three (of 300) custodians. We kept refining the search terms and adding custodians until we reached what proved in the end to be critical mass. By the time we had tested 5% of samples across 10% of the custodians, our keywords were on the money, as shown by the fact that search hit and review call-responsiveness rates remained steady, while instances of responsive documents in non-hit samples disappeared. Also, these results remained fairly constant throughout the review. While critical mass or the required comfort level in the selection criteria may vary from case to case, when one

Eugene Eames is a data-analytics consultant at SPi, a provider of electronic-discovery and litigation-support services. He can be reached at geames@spi-bpo.com.

needs to explain the process to judges or opposing counsel, the overall approach is easy to understand and provides a significant degree of comfort.

In another matter, we helped a different pharmaceutical company in a case that involved a product that affects the renal system. The legal team, of course, included *kidney* as a search term, but this resulted in an overbroad response because a large percentage of documents in a pharmaceutical collection may contain the word *kidney*, and those documents could have nothing to do with the issue at hand. However, by testing and refining the applicable search criteria, we were able to develop an accurate Boolean-based search string where only documents that included *kidney* were returned if other words were nearby (and certain others were not). The result was a much smaller corpus of documents to review, and we provided documentation and analytic test results to explain the more aggressive approach.

With a good data-analytics process, a provider can typically reduce the review set by an additional 20%. For example, if a traditional keyword-based approach results in a document set that is 40% of the original corpus, with data analytics this set might be reduced to 20%. Because review costs are such a high percentage of discovery expenses, the resultant savings can be enormous, but even more valuable might be the peace of mind that testing provides. The goal of data analytics is not just to reduce the review set, but also to target the correct review set.

A good data-analytics process not only refines keywords by testing and refining the hits that result in searches, but also samples and tests the non-hits to ensure that all responsive documents are getting into the review queue. And with the help of automated and manual feedback from review platforms and review teams, criteria-testing can continue as attorneys and paralegals review documents to make sure no effort is wasted, and that an appropriate percentage of documents is being marked as responsive. Using this approach takes advantage of the review team's work product and acquired knowledge by using those commodities to continually validate and possibly enhance the selection criteria. We believe that non-hit analysis and

feedback-loop strategies can be extremely valuable from a defensibility standpoint.

BENEFITS OF DATA ANALYTICS

The most important benefit of adding data analytics to an e-discovery project is that it makes the culling process much more accurate. The client can be confident that it is optimizing review by providing a set of documents most likely to be responsive. This can be accomplished by iteratively testing the documents returned, and the documents not returned, by selection criteria, and then modifying the selection criteria accordingly. The legal team can be comfortable knowing that it is, in fact, reviewing largely responsive documents, and not leaving behind possibly responsive documents that were inadvertently omitted from review because the electronic search tools were applied blindly and without validation of results.

Besides comfort, data analytics increases production defensibility. It's a common sense approach that makes sense to judges, who are sometimes uncomfortable with computer-assisted review. Because legal research tools such as Lexis and Westlaw have been so pervasive, most attorneys and judges are comfortable with the concepts of keyword and Boolean search. But that also means they have some understanding of the potential difficulties: They've used overbroad terms before, and they've used terms that they expected to return the proper document but the terms haven't. A simple system of sampling, testing and refining makes a lot of sense. And testing non-hits allays the nagging fear that responsive documents have slipped through holes in the net the teams have cast to snag relevant responsive documents.

A good data-analytics process is also well documented. For example, a document is created that shows the path from the brainstormed keywords to the final search criteria. A judge or opponent can see exactly what changed from search Iteration 3 to search Iteration 4, along with statistical results based on search-hit analysis. Also documented are results of the review, which enables teams to show that the results of samples held up — or to make changes if they did not hold up.

The most obvious benefit of data analytics, however, is cost-savings. In most cases,

the most expensive part of the discovery process is in the time that attorneys and paralegals spend reviewing documents. The less time legal professionals spend reviewing clearly non-responsive documents, the less expensive the overall review.

Lowering the overall cost plays out in two ways: 1) Better search criteria result in a higher percentage of responsive documents; and 2) the testing of the non-hits that really enables big savings. By reviewing a small percentage of non-hits, counsel can be comfortable in not reviewing the rest. The result is a much smaller review set, and a much more efficient, and less expensive, review.

CONCLUSION

While companies are always under pressure to reduce costs, data analytics is really a tool designed to get the correct set of documents. By doing so — in a defensible way — companies can reap the benefits of an optimized review while increasing confidence in their productions as a whole.



Reprinted with permission from the June 2007 edition of the LAW JOURNAL NEWSLETTERS - E-DISCOVERY LAW & STRATEGY. © 2007 ALM Properties, Inc. All rights reserved. Further duplication without permission is prohibited. For information, contact 212-545-6111 or visit www.almreprints.com. #055081-06-07-0002



11400 Burnet Road
Building 5, Suite 5110
Austin, TX 78758
Tel: 512 275 9595
E-mail: legal@spi-bpo.com