DrinkerBiddle

Publication - 10/01/2010

The Demise of Linear Review

Client Alert

By Bennett B. Borden

There is a familiar image from the 1920s of row upon row of bespectacled clerks, sporting green eye shades and clacking away on adding machines as they tracked and totaled the business of Wall Street. In the context of document review, a similar scene may still be observed; row upon row of attorneys clicking away in sunless rooms, reviewing the documents of corporate America. On Wall Street, the rows of clerks have been replaced by racks of servers. Today, with the emergence of advanced search and categorization technologies, a similar evolution in efficiency is happening in electronic discovery.

Technology has created a deluge of data

Computers have enabled the creation of the information economy – and unleashed an unprecedented deluge of data. On any given day, more information is sent via email than is contained in the entire print collection of the Library of Congress, and more data are created by American businesses than exist in every book in every library in America. As of 2002, there were 5 exabytes (5 billion gigabytes) of data stored in computers, the equivalent of every word ever spoken by human beings. As of 2009, there were about 988 exabytes of data, or roughly 64,220,000,000,000,000,000 pages of text. If printed, they would reach from the earth to the moon 75 times.[1]

The explosive growth in the volume of data can create a crippling financial and administrative burden on parties responding to discovery requests to identify, collect, review, and produce data.[2] Perhaps exacerbating this problem is the manner in which attorneys are accustomed to conducting discovery. An attorney 20 years ago who drafted a document request for "All documents concerning [anything]" would receive in response a more or less manageable amount of information. Today, the response would be inordinately different. We are accustomed to thinking of information accrued in the course of business as an asset of significant enterprise value. But, to the extent that a particular document is likely to be the object of a discovery request, it potentially can also represent a very real liability. The cost of collection, review and production often exceeds $2 per document – and corporations produce and store many billions of documents annually.

Paper processes in a digital world

Not long ago, complying with discovery requests often meant that young lawyers would wade through mountains of boxes filled with dusty, poorly organized documents. Faced with such a task, the only thing to be done was to read each document in serial fashion. This is the quintessential linear document review: tedious, tormenting, and terribly inefficient.

The introduction of computers and software applications that allow for "on-line" review replaced the flipping of pages with the somewhat more efficient clicking of a mouse. More useful still were the term searches that quickly became possible. Term searches could be used to help find relevant documents more quickly.

Unfortunately, once the term searches were executed, the data volumes that remained to be reviewed were still enormous and, for the most part, had to be coded through the old, inefficient, paper paradigm of document-by-document review. Term searches can reduce the data collection by a large fraction, but when one starts with tens – or hundreds – of millions of pages, what remains to be reviewed after the term-filtering reduction is still a significant burden to bear. It is reminiscent, if perhaps hyperbolic, of when King Henry VIII commuted the sentence of Anne Boleyn from burning to beheading. It was an improvement, but still ....

The evolution of search and categorization technologies in discovery

The challenge facing attorneys conducting a document review is similar to the one facing computer scientists designing a search engine: how does one retrieve all, or the vast majority of, the documents relevant to a query, while retrieving a minimal number of irrelevant ones? The nomenclature may be somewhat different – an attorney may refer to responsiveness rather than relevance – but the goal is the same. It is unsurprising, then, that a variety of search technologies, borrowed from the realm of computer science, have been brought to bear on electronic discovery.

The most widely utilized search technology is also the oldest: Boolean search.[3] In using common legal research tools such as Lexis® and Westlaw®, many of us have grown familiar with its logical operators (AND, OR, NOT, and the like). The fact that Boolean search has grown so popular, despite intrinsic limitations, can be attributed to its remarkable simplicity: if a particular search term appears in any document, then the document is returned as part of the result set. However, the superficial appeal and readily comprehensible nature of Boolean search serve to mask its flaws.

To understand these flaws, imagine going to the supermarket looking for something sweet. In order to find sweet things you ask for "any item that contains sugar." The problem here is obvious. An awfully large number of food items have sugar in them. Furthermore, you will find food items that are sweet, but that do not contain sugar, *e.g.*, those items that contain honey, high fructose corn syrup, or artificial sweeteners. Also, your shopping cart is likely to include many items that have only insignificant amounts of sugar but that don't taste sweet.

These are the key weaknesses of Boolean search. First, you must know in advance exactly what you are looking for, and then enumerate each possible lexical representation of it (in this case, every type of sweetener). Second, if you fail to enumerate each possibility, you inevitably will miss some sweet items (*e.g.*, items containing artificial sweetener). Third, assuming that you are able to enumerate each possibility, the results are likely to be over-inclusive (*e.g.*, items that contain only miniscule amounts of sugar).[4]

The flaws in Boolean, or first generation, search led to the development of second generation, or 2G, technologies in an effort to address them. One of these 2G technologies, called term frequency/inverse document frequency (or "TF/IDF"), was introduced in the mid 1980s in an attempt to make sense of documents identified by search terms by ranking them according to some sort of relevance. Most of us employ TF/IDF every day when we use an internet search engine. When we search for "products containing sugar," the hits are ranked in part by their relevance as established by their TF/IDF scores. This same kind of relevance ranking is used in some 2G document review platforms to rank documents for review.

TF/IDF judges a document's relevance through an analysis of how often a term appears within it (its "term frequency") in comparison to how often the same term appears in documents across the entire data set (its "document frequency"). If a term appears often in a document, it gets a high term frequency score. But if that same term appears in every document across the dataset (getting a high document frequency score), then it is difficult to imagine that any one document is more relevant than another, at least in relation to that term. Thus, under TF/IDF theory, the most relevant document in relation to a particular term is the one in which a term appears often (a high TF), but does not appear across the entire dataset (a low (or high inverse) DF). This was a marked improvement on Boolean search, in that now the documents within a dataset had some sense of relevance ranking instead of just one large set of documents with no differentiation among them. In terms of a document review, the dataset could now be addressed beginning with the most likely relevant (*i.e.*, responsive) documents.

Another 2G improvement over simple Boolean search is commonly described in e-discovery vernacular as "concept clustering." It attempts to group together documents containing terms that overlap to a similar degree and in a similar way. At its most simplistic, concept clustering is an integration of both Boolean and relevance ranking technologies. If two documents contain a term in a similar way, they should be considered together and presented as related (*i.e.*, in a cluster). To the degree that the cluster is meaningfully internally coherent, this improves review efficiency because in many cases a reviewer can consider and code the clustered documents as a block.

Another 2G search technology introduced the idea of synonymy. Boolean searches are limited by the terms applied against a dataset. Search for "sugar," for instance, and you won't get "honey." But a synonymous search engine will infer that, because the term "sugar" commonly appears closely associated with sweet items, the search should return documents containing other terms that also appear frequently in a sweet context (*e.g.*, honey). This helps to pull into the return set documents that may be relevant but that otherwise would have been overlooked.

All of these 2G search technologies are an improvement over simple Boolean searching. Synonymy helps resolve the under-inclusiveness problem of Boolean, and relevance ranking and concept clustering help resolve the over-inclusiveness problem (by identifying the most relevant documents to focus on first, and by grouping other relevant documents with them). And, these 2G technologies are increasingly being incorporated into commonly used document review applications.

2G search technologies, however, contain an inherent weakness: their dependence on the correct identification of search terms. With TF/IDF, a document's relevance is determined in relation to the search term. The same dependence holds true with clustering and synonymous search technologies. If the right terms are not applied against the dataset, responsive documents will almost certainly be missed (garbage in, garbage out). Another weakness is that 2G search technologies largely perceive documents as merely a collection of words, without gleaning information from where those words appear in a particular document, except in relation to other words. But, a document is not merely a collection of words, it is a collection of words that appear in a particular place within a document or, in the case of metadata, outside of the document.

The vast majority of documents contain similarities in structure. For instance, if a word appears in an email, it matters greatly where in the email the word appears (*i.e.*, in the to, from, cc, bcc or subject field, or in the body, header, or footer). The location of a word in a non-email document also matters, such as in the title, the header, the footer, or a paragraph, or within a data cell or presentation slide. And what about the history of a document? Information can be gleaned from when a document was created, saved, modified or printed and by whom. It also might matter that an email or document was grouped with others in a subfolder, or was created, modified or deleted just before or after a key event.

The most advanced third generation or 3G search technologies leverage all of these meaningful data points to compare and group documents. 3G technologies help alleviate 2G's dependence on correctly identifying search terms. They return not only documents that contain synonymous terms, but also documents that may not contain the term but that are closely associated with term-returned documents in other ways (*e.g.*, they were in the same subfolder, were created or deleted near in time, were distributed among the same group of people, and many other such characteristics). Leveraging all of the many data points now available about electronic documents, these 3G technologies group documents together. The degree of granularity and internal thematic cohesion within a category of documents is such that the documents often can be considered and evaluated together. It then becomes possible to browse a set of documents, find those groups that are relevant, and discard those that are not.

Intelligent use of new technologies can reduce the burden of review by an order of magnitude

In the e-discovery context, the application of 3G categorization technologies can drastically increase document review efficiency. Experiments with these technologies have produced some extraordinary results. Here are a few of them:

In one review of about 5,000 documents, a team of five reviewers used a common 2G review application. The documents were identified for collection by Boolean searches and then reviewed in a standard linear fashion (small groups of documents were assigned to each reviewer, who reviewed them one by one). The review required 110 working hours at a rate of about 45 documents per hour. This is a bit slower than the usual 50 to 60 documents per hour that is used as an e-discovery industry average, but the documents were fairly technical and required more than the average level of scrutiny. The same set of reviewers then reviewed 7,500 documents for the same matter identified in a second collection by Boolean searches. This time, the reviewers used a review application with 3G categorization capabilities. This second review, with substantially similar documents and identical coding protocols, was completed in 55 working hours at a rate of about 136 documents per hour.

In another experiment, two groups of reviewers of roughly equal experience were provided with the same set of about 60,000 documents. One group reviewed the dataset using a review application that employed 2G technologies while the other group used an application that employed 3G categorization techniques. The 3G reviewers completed the review *six times* faster than the 2G reviewers. Quality control testing also revealed that the 3G review set was coded significantly more accurately than the 2G review set.

In another example, a review was conducted using 3G categorization technologies, and based entirely upon non-linear review principles. Instead of identifying a set of documents and farming them out in smaller sets to a reviewer, small teams of reviewers were formed and each team was assigned a topic based upon discovery requests, and then tasked with finding responsive document. The reviewers were "unleashed" from linear review, and allowed to pursue a topic by categorization, custodian, date, search terms, and any other characteristic that they found worthy of pursuit. About 3 million documents were collected for review. An application of advanced categorization techniques resulted in the culling of about 1.5 million obviously irrelevant documents in about 10 hours using four senior reviewers. A group of 22 reviewers was then divided into three topic teams to review the remaining 1.5 million documents.

Using the non-linear review model, which the reviewers dubbed "rabbit hunting" or "choose your own adventure reviewing," the reviewers were set upon the dataset to pursue their topics however they chose. The 22 reviewers completed the review of 1.5 million documents in 124 hours (or roughly 15 working days). The daily average rate of the review was 318 documents per hour. While these are extraordinary results, some caveats should be noted. Even though the reviewers culled about half of the dataset prior to review, the original collection was wide in scope and contained a great deal of nonresponsive material (above 90 percent). And, because of the high coherence of the documents within each computer-generated category, large swaths of documents were able to be coded as a block, driving up the rate of documents reviewed per hour. However, even with these caveats, the review was extraordinarily fast, efficient, and accurate. A quality control analysis found an error rate of less than 3 percent, much lower than error rates using other review protocols.

A post-review analysis of the review metrics and a debriefing of the reviewers revealed some of the reasons for the great efficiencies achieved using advanced 3G categorization techniques coupled with non-linear review. By focusing on topics related to the document requests, and using categories and other document characteristics, reviewers were able to identify documents that were highly cohesive in relevance. That, in turn, enabled them to make more accurate coding decisions by staying on the same topic instead of jumping from one topic to another, as is often the case when presented with a somewhat random set of documents to review linearly.

They also were able to pursue investigative themes. For example, several reviewers indicated that they would pull up a return set by custodian, category, or date range, for instance, and begin to "poke through" them. They would then "come upon a scent" of a potentially relevant fact or circumstance and would then pursue it. They were free to back out of a return set and pursue the theme through another avenue (*i.e.*, tracing a thread of emails, using a search term revealed in a document, or looking at particular custodians' documents around a certain time when a meeting was to occur). This not only led to more accurate coding, it very quickly developed key facts that we used to generate discovery requests, interrogatories, requests for admissions, and deposition preparation. Also, the reviewers indicated that they enjoyed the review to a much greater degree than any other that they had conducted because they were "engaged in the hunt" instead of mind-numbingly clicking away document after boring document. No doubt this also added to the speed and accuracy of the review.

The most potent effects are achieved by combining powerful technologies with legal skill

By skillfully using advanced technologies, there is a way to avoid the painfully inefficient and expensive linear review models that have predominated in the past. These same technologies also can be used to identify and collect a more targeted and cohesive group of documents, further reducing inefficient effort. With the increasing adoption of better technological tools and more efficient practices, we may be seeing the demise of wasteful linear review. In its place, we will see a more cost-effective, focused, and defensible review process driven by smart search technologies – and by attorneys who can skillfully leverage them.

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

Bennett B. Borden is a partner at Williams Mullen and chairs its Electronic Discovery and Information Governance Section. A litigation attorney, Mr. Borden focuses his practice on Electronic Discovery and Information Law. Mr. Borden is a member of the steering committee for the Electronic Discovery Section of the District of Columbia Bar, and a member of the Science Law and Technology Section of the American Bar Association. He is a guest lecturer on E-Discovery at the Georgetown University Law Center and a member of Working Group I on Electronic Document Retention and Production of The Sedona Conference.

[1] Jason R. Baron and Ralph C. Losey, *Electronic Discovery: Did you know?*, available at http://www.youtube.com/watch?v=bWbJWcsPp1M

[2] George L. Paul and Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 RICH. J.L. & TECH. 10 (2007), http://law.richmond.edu/jolt/v13i3/article10.pdf.

[3] "Boolean" refers to the logical calculus devised in 1847 by English mathematician George Boole, while developing the precursor to the modern computer that he called an analytical engine. It is based upon binary values of true or false, which developed into the zeroes and ones of modern computer programming.

[4] An interesting overview of the weaknesses of term searching and the legal industry's early attempts at using more advanced search technologies appears in, *The Sedona Conference® Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery*, 8 The Sedona Conference J. 189 (2007).