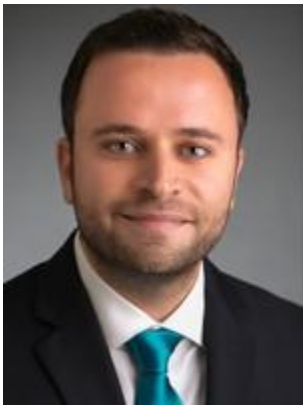
Portfolio Media. Inc. | 860 Broadway, 6th Floor | New York, NY 10003 | www.law360.com
Phone: +1 646 783 7100 | Fax: +1 646 783 7161 | customerservice@law360.com

# Making The Most Of Document Analytics

*Law360, New York (December 1, 2015, 10:34 AM ET) --*



Rand Ghayad

The volume of documents routinely subject to discovery poses challenges in investigations and litigation that extend beyond e-discovery. While predictive coding is gaining increased acceptance as a procedure for identifying responsive documents with less manual review, there is less appreciation of how document analytics can add value in answering document related research questions, or otherwise helping to identify and analyze documents in ways not practical with keywords alone. Having reduced reliance on manual document review to decide which documents to produce, the challenge is to determine quickly what the documents reveal about the critical issues in the case.

Document analytics offer large potential payoffs in the conduct of investigations and case development. An advantage of using computer programs (i.e., algorithms) to analyze documents is that, unlike manual review, algorithms can be run across all documents in the universe at relatively limited cost. While the results of computerized document classification may not be perfect, analyzing all documents collectively reveals patterns not visible from targeted manual review. For example, important patterns of communication concerning particular topics may only become apparent once all messages are analyzed and mapped. Furthermore, algorithms can be used to gather individual pieces of similar information of interest across an entire database, for example pricing information, providing a basis for economic analysis that would otherwise be far more cumbersome to perform.

Predictive coding is an example of a document analytics procedure that has now been accepted by a number of courts. This is an important development not only for increasing the efficiency and lowering costs of discovery, but because it has established the use of sophisticated document analytics as an acceptable approach for document-intensive investigations. This poses a challenge for law firms to bridge the gap between the technology platforms available to access documents, and the useful lines of investigation that could be conducted using sophisticated document analytics.

In this article, we summarize court opinions on the superiority of using document analytic methods, in particular predictive coding, over keyword searches; describe how predictive coding works; and provide an illustration of how a closely related method, topic modeling, can be used in document intensive investigations.

**Several Courts Have Concluded That Predictive Coding Can Be Superior to Keyword Searches**

In many cases, keyword searches can be overinclusive. That is, they return responsive documents with an overwhelming set of irrelevant documents. They can also be underinclusive. For example, the lack of standardized terms used in conversations and documents makes it hard to retrieve all documents relevant to a given set of search terms. Searching for the words "automobile" and "car" will miss references to "BMW" and "Mercedes." The mere formulation of a query or keywords is difficult if the information being targeted can be described in several different ways. Moreover, simple search queries may return ambiguous uses of the keywords being searched. It may retrieve "hits" of the words that are not really relevant to an inquiry. And, of course, keyword searches generally will not retrieve any documents containing a keyword that is misspelled, either in the query or in the documents.

For these reasons, keyword searches have been criticized in several recent decisions. In the landmark decision in National Day Laborer Organizing Network v. U.S.Immigration and Customs Enforcement Agency, the court questioned the general effectiveness of keyword searches. The court examined the reasonableness of various government agencies' search efforts in response to a Freedom of Information Act request and asserted that "simple keyword searching is often not enough … There is increasingly strong evidence that '[k]eyword search[ing] is not nearly as effective at identifying relevant information as many lawyers would like to believe.'"[1] Similarly, in Disability Rights Council of Greater Wash. v. Wash. Metro Transit Auth, the court suggested the parties consider using "concept searching, as opposed to keyword searching, [as it is]… more efficient and more likely to product the most comprehensive results."[2]

In Da Silva Moore v. Publicis Groupe, the first validation of the use of predictive coding by a U.S. court, the parties agreed to identify responsive documents from among a universe of over three million documents based on the review of a small sample.[3] The judge reminded them that technology assisted review "works better than most of the alternatives, if not all of the [present] alternatives. So the idea is not to make this perfect, it's not going to be perfect. The idea is to make it significantly better than the alternatives without nearly as much cost." The court concluded that predictive coding "now can be considered judicially-approved for use in appropriate cases."[4]

In EORHB Inc. et al. v. HOA Holdings LLC, the Delaware Chancery Court essentially required the parties to use predictive coding to meet their document production obligations in a pending matter, prior to either party actually proposing its use.

In 2014, Dynamo Holdings v. Commissioner of Internal Revenue followed in the footsteps of Da Silva Moore. In this case, predictive coding saved Dynamo months in e-discovery and over half a million dollars in document review expenses.[5] The court embraced the fact that "the technology industry now considers predictive coding to be widely accepted for limiting e-discovery to relevant documents and effecting discovery of [electronic stored information] without an undue burden."[6] The court further asserted that "although predictive coding is a relatively new technique … e-discovery and electronic media has advanced significantly in the last few years, thus making predictive coding more acceptable in the technology industry than it may have previously been."[7]

Most recently, in the 2015 court case Rio Tinto v. Vale SA, the court encouraged the use of predictive coding where appropriate.[8] The parties agreed to use predictive coding and the court approved, stating that "in the three years since Da Silva Moore, the case law has developed to the point that it is now black letter law that where the producing party wants to utilize [technology assisted review] for document review, courts will permit it."[9] Predictive coding has also been endorsed in regulatory

proceedings. In the proposed merger of Anheuser-Busch InBev and Grupo Modelo, for example, the U.S. Department of Justice approved a request to use predictive coding to review documents related to the antitrust review of the proposed merger.[10]

**How Predictive Coding Works**

The predictive coding model used in Da Silva Moore, was based on the review of a random sample of 2,399 emails from the entire database of over 3 million. The sample emails were classified as responsive or nonresponsive and were used as a "seed set" to train the predictive coding software. The trained software was then used to predict the coding for the larger universe of documents. The parties discussed the optimal number of iterative rounds to stabilize the software's training. Between rounds of training the software and the model were recalibrated, to correct miscoding. After the seventh round, all the documents in the database were coded and the defendant reviewed a random sample of the discards (i.e., the nonresponsive documents) to verify that the software did not set aside documents that were, in fact, highly relevant.[11]

In contrast to traditional keyword searching based on specific words or phrases, concept searching is a more sophisticated approach that does not require the parties to agree on and identify all possible keywords of interest up-front. Predictive coding is a form of concept searching that can classify documents based on concept similarity, even if all the target words are not contained in the document.

Underlying typical predictive coding models are scoring systems that assign weights to words, phrases and metadata in the training set, essentially converting text to data. Scores (or weights) reflect the information content of each data element — a very positive weight makes a doc likely relevant, while a very negative weight makes a document likely irrelevant. The initial weights are then calibrated by testing them against each document in the sample by iterative trial and error. The model scores each document based on the weights assigned to the words it contains. Just as human reviewers reach different decisions on the relevance of the same document, a predictive coding model may make predictions that do not match an attorney's decisions in every instance. Documents that are misclassified by the predictive coding model are used to adjust the model weights during the training rounds. The training rounds end when the model predictions have reached stability, or are sufficiently accurate, balancing the cost of missing responsive documents against the cost of including non-responsive documents.

**Topic Modeling in Investigations**

Predictive coding is an example of a document analytic method that involves "supervised machine learning." In particular, the algorithm learns from human decisions and then applies those decisions to new data. It is considered supervised because it uses a training set of manually reviewed documents selected and tagged by the user. The selection of the training set, the settings of the supervised learning algorithm, and the way in which it is actually relied upon, are important considerations in supervised learning.

In contrast, algorithms that can analyze data without supervision (i.e., without a training set) are sometimes used for "early case assessment" purposes. These algorithms are considered unsupervised because the system derives the themes without specific human intervention. The system seeks to discern patterns already inherent within the data. An example is the derivation of "topics" from within litigation document collections. This technique is referred to as "topic modeling." The model uses the relative frequency of words, phrases or metadata to group similar documents together in clusters

without the need for manual review. The topics are defined by the combination of words, phrases and metadata that appear together most often. Topics are commonly represented as word clouds. Word clouds can also be labeled based on manual review and characterization of the word combinations produced by the procedure.

**Implications for Case Work**

Recent case law reflects a clear progression towards judicial acceptance of document analytics. Those courts and regulators that have embraced the use of document analytics have noted that more traditional tools of document review, such as manual reviews and keyword searches, can simply be too expensive and ineffective in an era of big data. With so many documents and so much data subject to discovery, determining how to best use standard data science tools and customized algorithms in investigations will increasingly become a key to efficiency and success.

—By Rand Ghayad, Paul Hinton, Mark Sarro and Michael Cragg, The Brattle Group Inc., and David Cohen, Reed Smith

*Rand Ghayad, Ph.D., is an associate in The Brattle Group's Cambridge, Massachusetts, office. Paul Hinton is a principal in The Brattle Group's New York office. Mark Sarro, Ph.D., and Michael Cragg, Ph.D., are principals in the firm's Cambridge office. David Cohen is a partner in Reed Smith's Pittsburgh office and leader of the firm's global records and e-discovery practice group.*



[1] Nat'l Day Laborer Org. Network v. United States Immigration & Customs Enforcement Agency, 877 F. Supp. 2d 87, 2012 U.S. Dist. LEXIS 97863, 2012 WL 2878130 (S.D.N.Y. 2012), citing Maura R. Grossman & Terry Sweeney, What Lawyers Need to Know About Search Tools: The Alternatives to Keyword Searching Include Linguistic and Mathematical Models for Concept Searching, Nat. L. J. (Aug. 23, 2010).

[2] Disability Rights Council of Greater Wash. v. Wash. Metro. Transit Auth., 242 F.R.D. 139, 2007 U.S. Dist. LEXIS 39605 (D.D.C. 2007).

[3] Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182, 2012 U.S. Dist. LEXIS 23350, 18 Wage & Hour Cas. 2d (BNA) 1479, 2012 WL 607412 (S.D.N.Y. 2012)

[4] Da Silva Moore.

[5] Dynamo Holdings L.P. v. Comm'r, 2014 U.S. Tax Ct. LEXIS 40, 143 T.C. 183, 143 T.C. No. 9 (T.C. 2014).

[6] Dynamo Holdings.

[7] Dynamo Holdings.

[8] Rio Tinto PLC v. Vale S.A., 2014 U.S. Dist. LEXIS 174336 (S.D.N.Y. Dec. 17, 2014)

[9] Rio Tinto PLC v. Vale S.A., 306 F.R.D. 125, 2015 U.S. Dist. LEXIS 24996, 2015 WL 872294 (S.D.N.Y. 2015)

[10] Joe Palazzolo, “Software: The Attorney Who Is Always on the Job,” The Wall Street Journal, May 6, 2013, available at http://www.wsj.com/articles/SB10001424127887324582004578460860324234712 (last accessed November 16, 2015).

[11] Da Silva Moore.

---