Discovia™
Managed eDiscovery

Discovia Manages 2.1 Million Documents in 30 Days, Saves Client More than $2 Million with Defensible Culling, Predictive Coding, Keyword / Domain Search

Case Study | **Client:** Law firm representing an energy company | **Project Team:** Discovia eDiscovery experts

THE PROBLEM

An energy company was heading into arbitration in a large-scale construction dispute. The law firm representing the company needed to review and produce responsive documents from a data set of 2.1 million documents in a 30 day time period. A cost effective and efficient solution to cull and review the documents for relevance was needed to manage this large data set in the tight timeframe required.

THE SOLUTION

Working closely with Discovia's consultants, the law firm decided on a workflow that would include the use of Discovia's proprietary Intelligent Culling Application (ICA), Equivio Zoom's Relevance predictive coding module, and keyword search.

STEP 1: Junk File Analysis and Culling

Using file extension sampling and exclusion, 112,285 documents were excluded as computer generated files or image files. Additionally, domain parsing was used to identify 230 domains and 57,000 additional documents as "junk" mail. Because Equivio Relevance's predictive coding technology was selected as the second phase of the culling process, the client opted not to perform any key-word filtering at this point to further reduce the data set.

STEP 2: Equivio Relevance Powered by Discovia

Once initial culling was complete, Discovia's team of technology assisted review consultants worked with the client to configure the project for Relevance. Files not suitable for Relevance (such as unusually large files and files with no extracted text) were set aside and loaded into Relativity for standard review. The remaining files (approximately 1.6 million documents) were loaded into Relevance for technology assisted review. Because Relevance only requires the text of the documents, not complete native files or images, the loading process is very fast. In this case the files were loaded in a matter of hours, which served the case team well in this time-sensitive matter.

The client selected a case expert who was well-versed in the case data and able to definitively distinguish relevant from non-relevant documents to train the system. Additionally, in order to comply with the agreed upon discovery protocol in the matter, the client disclosed their plans to use technology assisted review to the opposing party.

The Assessment Phase

The case expert was provided a random set of documents to begin the "assessment phase" of the Relevance workflow. The purpose of the assessment phase is to provide a control set of documents to estimate the richness level of the data (aka the percentage of relevant documents in the document universe) and ultimately determine if the data is rich enough for the machine learning to work. For best results, a richness level of at least 10% is recommended for Relevance to produce optimum results. The assessment phase typically requires the case expert to review ~500 documents. In this case the expert reviewed 520 documents and the system determined the richness of the document population to be 24.5% - well above the minimum recommended for best results.

The Training Phase

With the assessment phase complete, the expert reviewer was able to move on to the interactive machine learning (aka training) phase of the Relevance workflow. During this phase, the case expert reviewed documents in batches of 40 to train the system on the issue of relevance. The expert marked each document in the batch relevant or non-relevant, allowing the system to learn what constitutes relevance. Using the information gathered from each batch, the system presented the expert with a new batch of documents from which it could gain further insight into the nuances of relevance. When the system had learned all that it could such

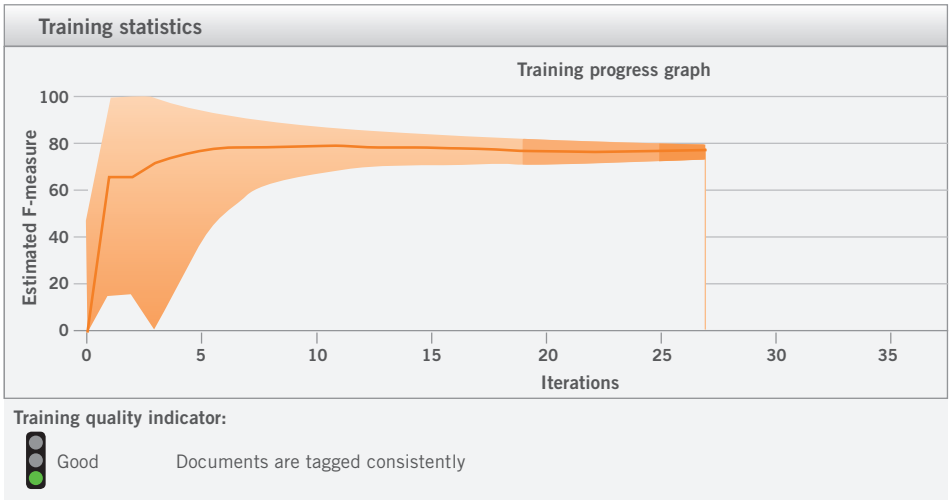


Figure 1

100 California Street, Suite 800 San Francisco, CA 94111 p415.392.2900 f415.392.2902 discovia.com

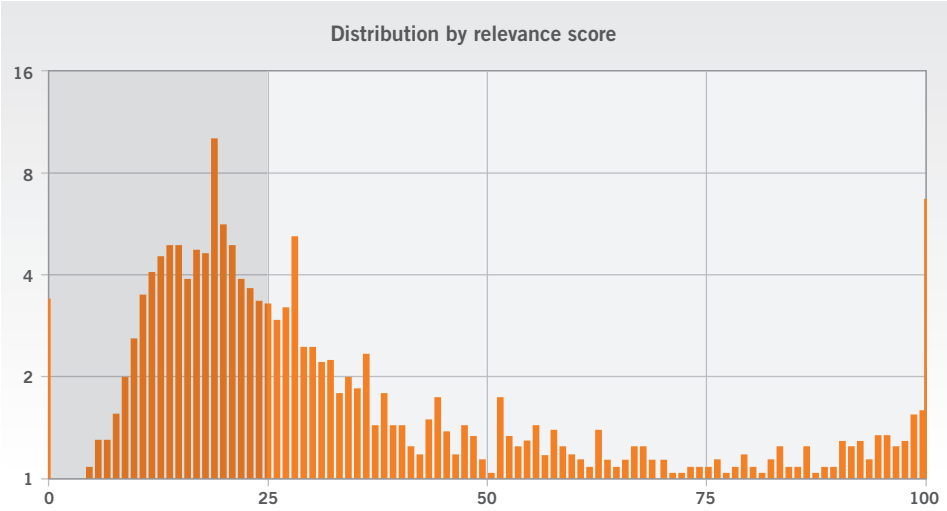


Figure 2

that the expert reviewing another batch of documents would not yield additional information, the system indicated that "stabilization" had been reached.

In this matter, the expert had reviewed 27 batches (only 1080 documents) when the system reached stability. Figure 1 is a screenshot from the Relevance interface that shows that the system reached stabilization.

The Batch Calculation Phase

With training complete, the system could now appy a relevance score to every document in entire population. This process, called batch calculation, assigns a relevance score between 1-100 to every document.. Figure 2 shows the distribution of documents based on relevance score.

A Bump in the Road

After the batch calculation was run, the case team hit a bump in the road. They identified an additional data source that had not been collected that was likely to contain documents relevant to the matter. That data would now need to be incorporated into the Relevance workflow, assigned a relevance score and incorporated into their decision-making process for next steps.

Fortunately, the Relevance workflow is designed to easily incorporate additional data sets not contemplated during the interactive training phase. In this case, given that the nature of the new data was similar to the data that the system had already been trained on, the expert only needed to review one batch of 40 documents from the new data set to allow the system to recalculate the relevance scores for the entire document universe.

The Decision Phase

After the new data set was incorporated into the batch calculation, the case team set about making a decision on which documents would go through review. Working with Discovia's experienced consulting team, the case team used Equivio's decision matrix to determine a cutoff point for document production. The team decided that they would select a relevance cutoff point that provided them with a reasonable recall percentage. Those documents with relevance scores above the cutoff point would be promoted to Relativity for a privilege review. All non-privileged documents would then be produced en mass to opposing counsel.

The decision matrix that the system created for the project recommended a review cutoff point at 84.3% recall (32% of the data set). The matrix also showed the approximate cost of human review at that cutoff point, and the cost to find the next relevant document. Based on the needs and budget for the case, the case team has the ability to adjust the recall and review percentages to get a review cutoff and cost that is right for their matter.

Recall is a metric that is used to measure the number of relevant documents retrieved out of the total number of relevant documents in the document set. For example, if recall is 80%, that means that 8 out of 10 relevant documents are retrieved. For most projects, 100% recall is not an option. Typically, recall in the 80% range is considered very strong. To provide some perspective, an often-cited study* about recall using iterative keyword search showed recall to be only 20%. Another study that compared human review to technology assisted review showed recall from document by document manual review to be only 59.3%**.

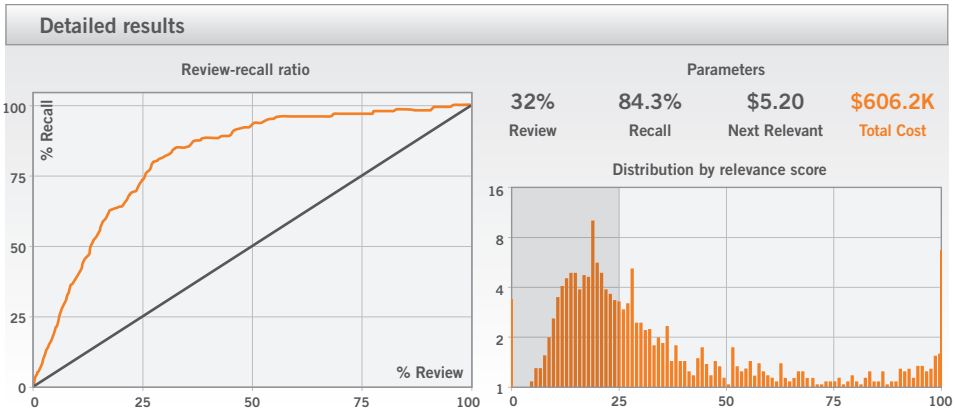
*David C. Blair & M.E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*, 28 COMMC'NS ACM 289 (1985)

** Bruce Hedin et al., Overview of the TREC 2009 Legal Track, in NIST SPECIAL PUBLICATION: SP 500-278, THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2009)

In this case, the team decided to go with Equivio's recommendation and move forward with promoting 32% of the data to Relativity for privilege review.

STEP 3: Privilege Search

The 32% of documents most likely to be responsive were loaded into a Relativity database hosted by Discovia for privilege searching and review. The case team



100 California Street, Suite 800 San Francisco, CA 94111 p415.392.2900 f415.392.2902 discovia.com

worked with Discovia's search consultants to craft privilege searches that would capture potentially privileged data based on email domain, attorney names, and keywords. These searches were used to prioritize the privilege review and focus reviewer's attention on the documents most likely to be privileged. Of the ~608,000 documents promoted to review, the potentially privileged search narrowed the review set to only ~134,000.

STEP 4: Quality Control

Sample documents that did not meet the relevance cutoff were reviewed to ensure that they did not included relevant material. A similar sampling was done on documents outside the potentially privileged search hits to ensure that the searches did not miss potentially privileged material.

STEP 5: Production

The client made two productions of documents to the opposing party, in compliance with the discovery protocol for the matter. Ultimately, ~485,000 documents were produced. A clawback provision that was part of the agreed-upon discovery protocol could be used in the event that any privileged material was inadvertently produced.

THE RESULTS

Using a combination of culling, predictive coding, and keyword and domain search, Discovia assisted the client with a cost effective and defensible approach to managing a data set of 2.1 million documents. After a case expert reviewed 1080 documents, a relevance score was assigned to the rest of the document population. After a privilege review that was expedited with the use of a robust set of potentially privilege searches, ~485,000 documents were produced to opposing counsel. Not only did the client meet the accelerated deadline set out in the discovery protocol, they did so in a manner that was far more cost effective than a more traditional approach of using keyword search for culling and linear review for responsiveness. They experienced a more accurate result and better recall of responsive data, while at the same time saving over $2,000,000 compared to traditional keyword search and linear review.

	Traditional Keyword Culling and Linear Review	Predictive Coding using Equivio Relevance
Contract Review	1.47M docs = $2.328M	134,000 docs = $212,000
Expert Review	NA	1600 docs = $13,300
Predictive Coding Technology	NA	$95,000
TOTAL	**$2,328,000**	**$320,300**
TOTAL SAVINGS = $2,007,700		

Assumptions:

Contract Review Cost = $95/hour (assumes a hybrid approach using contract and law firm attorneys)
Review Speed = 60 documents/hour
Expert Review Cost = $500/hour
Keyword Culling data reduction = 30%

100 California Street, Suite 800 San Francisco, CA 94111 p415.392.2900 f415.392.2902 discovia.com