Smarter Not Harder:
How One SME Applied Analytics to Organize a Document Review

Managed Review – Case Study & Cost Analysis

Innovative Discovery, LLC

Purpose

The purpose of this case study is to give a high-level summary of the time and cost savings that can be achieved by utilizing analytics to organize a review. This review was an internal investigation requested by a corporate client and needed to be completed in five (5) business days. Due to the sensitive nature of the investigation, a single in-house resource (Subject Matter Expert or "SME") was assigned to this initial review to assess whether there was an issue warranting further action.

Given the aggressive deadline the client was facing and the restriction of one SME able to review the documents, a linear review of 5,000 documents could not be completed in 5 business days. **Innovative Discovery, LLC ("ID") proposed a solution employing the use of analytics to strategically organize the review and as a result, the SME reviewed only 815 documents and the review was completed in three days, saving the client a total of $31,505.00.**

Documents for Review

1. Search terms were used to identify the review universe ("Relevant Search Terms").
   a. One of the search terms was a common term that would result in a large number of false positives. However, the term had to be included based on its context in the matter.
2. The Relevant Search Terms resulted in approximately 5,000 documents (documents with a Relevant Search Term hit and their family members) ("Potential Review Universe").
3. The Potential Review Universe was drawn from an email collection (emails and attachments).
4. The Potential Review Universe of 5,000 documents is approximately 2.3 GB (2,200 documents per GB).

Cost Savings – Linear Review vs. Review with Analytics

Review Category	Documents per Hour	Documents Reviewed	Hours to Complete Review	Cost to Complete Review
SME Review - Review With Analytics	40	815	20.4	-
Cost of Analytics	N/A	N/A	2.3 GB	-
TOTAL - REVIEW WITH ANALYTICS				**$ 8,495.00**
SME Review - Linear	50	5,000	100.0	-
TOTAL - LINEAR REVIEW				**$ 40,000.00**
****SAVINGS USING ANALYTICS**				**$ 31,505.00**

Summary of Review Workflow Using Analytics

1. Random Sample

A random sample of approximately 95 documents (95% confidence, 10% margin of error) was drawn from the Potential Review Universe and reviewed by the SME.

- Based on this review, 5% of the Potential Review Universe was estimated to be relevant, which is the equivalent of 250 documents. In other words, the Potential Review Universe has a 5% richness of relevant material, +/- a margin of error of 10%.

2. Email Threading & Identifying Inclusive Emails

ID used Relativity Analytics Email Threading to email thread[1] the Potential Review Universe.

- As part of the email threading process, the analytics engine identified the "inclusive" emails,[2] which are emails that contain unique content and must be reviewed ("Inclusive Emails").
- 50% of the Potential Review Universe was identified as inclusive emails, which is the equivalent of 2,500 documents.
- NOTE - Due to additional time and attention required to review inclusive emails, the average Documents per Hour review rate was reduced from 50 documents/hour to 40 documents/hour.

3. Clustering - ID then used clustering[3] to organize the Inclusive Emails for review.

4. SME Samples the Clusters

Since the Inclusive Emails were organized in clusters of conceptually similar documents, the SME started the review by sampling a cluster.

- If the sampled documents in the cluster were not relevant, then the SME moved on to the next cluster. The sampled cluster was tagged as "Non-Relevant Cluster."
- If the sampled documents in the cluster were relevant, then the SME reviewed all Inclusive Emails in that cluster.

At the completion of this review, the SME reviewed 25% of the Inclusive Emails using the logic above, which is the equivalent of 625 documents.

5. Coding

The SME coded the Inclusive Emails with one of the following tags:

- Wholly Relevant - This tag was used for Inclusive Emails that were relevant and all of the lesser-included emails were also relevant.
- Wholly Non-Relevant - This tag was used for Inclusive Emails that were non-relevant and all of the lesser-included emails were also non-relevant.
- Mix of Relevant & Non-Relevant - This tag was used for when there was a mix of relevant and non-relevant chains in the Inclusive Email

The coding applied to the Inclusive Emails was propagated to the entire email thread. Attachments to Inclusive Emails were coded on a document-level.

6. Validation Test

A random sample of approximately 95 documents (95% confidence, 10% margin of error) was drawn from the clusters that were sampled by the SMEs and tagged as "Non-Relevant Cluster."

- The SME reviewed this sample to see if any relevant material was missed. There were no relevant documents identified in the sample.

7. Review Results

This review was completed in three (3) business days and a total of only 815 documents were reviewed by the SME.

- 11% of the documents in the Potential Review Universe were tagged as "Wholly Relevant" or "Mix of Relevant & Non-Relevant."
- 11%[4] richness is within the 10% margin of error of the 5% richness resulting from the initial Random Sample Review.

[1] An email thread is a single email conversation that starts with an original email (the beginning of the conversation) and includes all of the subsequent replies and forwards pertaining to that original email.

[2] An inclusive email is an email that contains unique content not included in any other email, and thus, must be reviewed. An email with no replies or forwards is by definition inclusive. The last email in a threaded conversation is also by definition inclusive. Emails with attachments that were dropped once the email was replied to are also inclusives since they contain the attachments.

[3] Clustering is an organizational method whereby documents are segregated into mutually exclusive groups, or "clusters," of conceptually similar documents based on similar text patterns within the documents.

[4] Please note that since one of the tags is "Mix of Relevant & Non-Relevant," the 11% richness is a slightly inflated number.