Relativity®

White Paper

Measuring and Validating the Effectiveness of Relativity Assisted Review

Dr. David Grossman, Ph.D.

Contents

Introduction

In this white paper, we briefly describe a study conducted to evaluate the effectiveness of the Relativity Assisted Review workflow. We'll outline the core goals of the study, provide some background on it, detail the methodology we used, and finally cover the results.

Goal of the Study

Our goal was to evaluate kCura's Relativity Assisted Review workflow. In particular, we measured the effectiveness of the different types of rounds used in the workflow (e.g., training or quality control). Additionally, we tested the effectiveness of sampling as used in the workflow, to ensure that random samples are indeed representative of the full document population and that categorized samples are also representative of the full categorized document population.

Essentially, we answered three questions:

Q1: Does sampling work to estimate the total responsive and non-responsive documents in a collection?

Q2: Does Assisted Review's reporting functionality accurately reflect the true number of defects in the document universe? In other words, once categorization is completed, how well does a sample reflect the accuracy of the categorization that was just completed?

Q3: Does the Assisted Review process improve effectiveness?

Background on Computer-Assisted Review

The field of information retrieval has defined two core metrics for assessing the effectiveness of a search or document categorization tool. The first is **precision**, or the ratio of responsive documents in a collection to those responsive documents retrieved. For example, if we retrieve 100 documents and five of them are good, we have 5 percent precision. This might sound bad, but it doesn't tell the whole story. Suppose there were only five documents to find in the whole collection. As a result, we need another measurement: **recall**. Recall is the ratio of responsive documents to the total number of responsive documents in the full collection. If there were only five documents worth reading in the whole collection, and we had a system that found all five of them, recall would be 100 percent.

Assisted Review introduces another related metric called **overturn rate**. An overturn occurs when the category in which Assisted Review places a document is changed or overturned by a subject matter expert upon review, i.e. when a responsive document is found to be non-responsive or vice versa. Sampling is done to determine when the workflow should end, providing data on the effectiveness of the workflow at a current point in time.

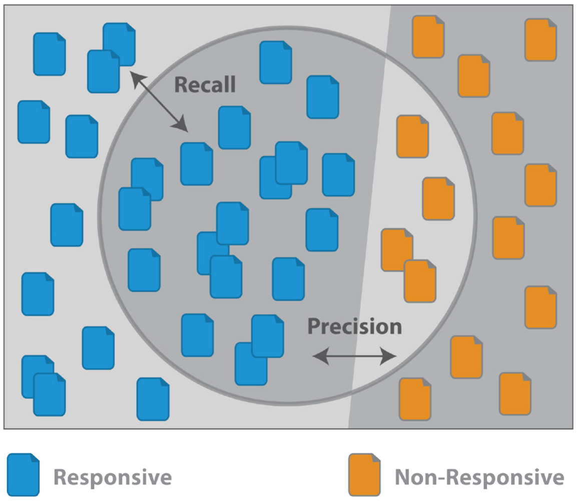

Figure 1: Recall and Precision
All retrieved items are within the circle, and the relevant retrieved items are the blue documents within the circle. Precision is the ratio of these relevant retrieved documents to all of the documents in this circle, both blue and orange. Recall is the ratio of the relevant retrieved documents to all of the blue documents in the chart, i.e. all relevant items.

Let's say we have a new sample of 1,000 documents, of which 100 are deemed responsive. If we review them and find out 10 documents are non-responsive, we would deem the overturn rate 10 percent. Note that the overturn rate essentially provides an updated value of precision and recall, because the numerator in both ratios is the number of correctly labeled documents. Hence, 90 of the 1,000 documents are now responsive, instead of 100, so precision drops accordingly. Overturns on non-responsive documents are simply indicators of improvements to precision and recall, while overturns on responsive documents are directly tied to reductions in precision and recall. ***Essentially, overturn is just an easier-to-understand term for ratios such as precision and recall.***

Methodology for Validating the Workflow

We took five established topics (201, 202, 204, 205, and 206) from the Enron 2009 TREC legal track and treated them as a single legal inquiry. To form a document collection, we took all of the documents that were manually assessed for those five topics. This was done to ensure that all documents we used contained ground truth judgments on relevance. This gave us a full collection of about 20,000 documents.

We ran the Assisted Review workflow on this collection five different times—each time starting with a different random sample of documents. For each of the five runs, we began with a series of two training rounds. Subsequently, we ran rounds with samples pulled from responsive documents. Finally, non-responsive documents were used for the samples in the last rounds. After each round, results were compared with a control set of categorized documents to see how well a control set can serve as a predictor of categorization accuracy for a full collection.

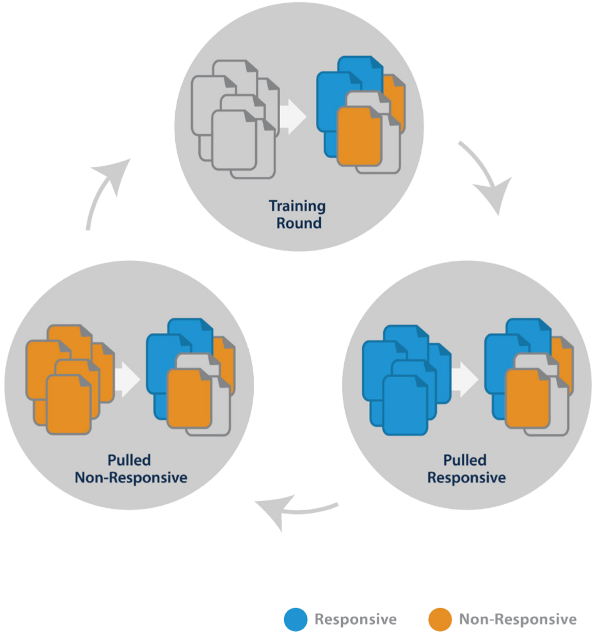

Figure 2: Workflow for Validating Assisted Review

A singular value decomposition of the full collection was done to allow for a more concept-oriented search using latent semantic indexing (LSI) along 100 dimensions. Essentially, the exemplars for a given category are turned into an LSI-based, or concept-based, query, using a text analytics product called CAAT from Content Analyst—such that a query exists for responsive documents and another query for non-responsive documents. The LSI query returns a ranked list of documents that contain a score that ranges from zero to one. Documents that are below a threshold of 0.5 are deemed not-categorized. Each time new exemplars are found, the representative query is updated. All random samples that were taken were selected with a sample size intended to ensure a 95 percent confidence level with a 2.5 percent margin of error.

Results

Our results address each of the three key questions:

Q1: Does sampling work to estimate the total responsive and non-responsive documents in a collection?

We sampled the collection five different times and compared the proportion of responsive and non-responsive documents in the sample to the proportion in the full collection.

Figure 3 displays the results, and it can be seen that the document makeup of both the full population and sample documents is very similar. The average difference in proportion between the full collection and each of the five samples was 0.45 percent. As the error rate was within our margin of error, statistical sampling was effective in estimating the proportion of responsive and non-responsive documents in the full collection.

Q2: Does Assisted Review's reporting functionality accurately reflect the true number of defects in the document universe? In other words, once categorization is completed, how well does a sample reflect the accuracy of the categorization that was just completed?

We sampled the population after categorization five times. This was done for two rounds of responsive document categorization. Figure 4 below shows a graph of the five trials. The average error was 0.84 percent.

Figure 4: Percent Error for Five Assisted Review Trials

	S1	S2	S3	S4	S5
Relativity Estimate Responsive	5,020	4,995	4,997	4,972	4,909
% Error	2.8%	1.6%	1.6%	0.87%	0.16%
Relativity Estimate Non-Responsive	14,235	14,267	14,273	14,308	14,075
% Error	1.0%	0.57%	0.54%	0.30%	2.23%

Q3: Does the Assisted Review process improve effectiveness?

For responsive documents, the average responsiveness precision in round one started at 59.5 percent and improved to 79.49 percent. Responsive recall started at 70.26 percent and improved to 84.01 percent. For non-responsive documents, the average precision started at 86.93 percent and improved to 93.87 percent. The average recall started at 60.54 percent and improved to 68.38 percent.

Figure 5 shows the precision and recall as the Assisted Review workflow progressed. Since the goal is to find as many relevant documents as possible, we believe "responsive recall" is the key metric. This metric clearly improved as the workflow progressed.

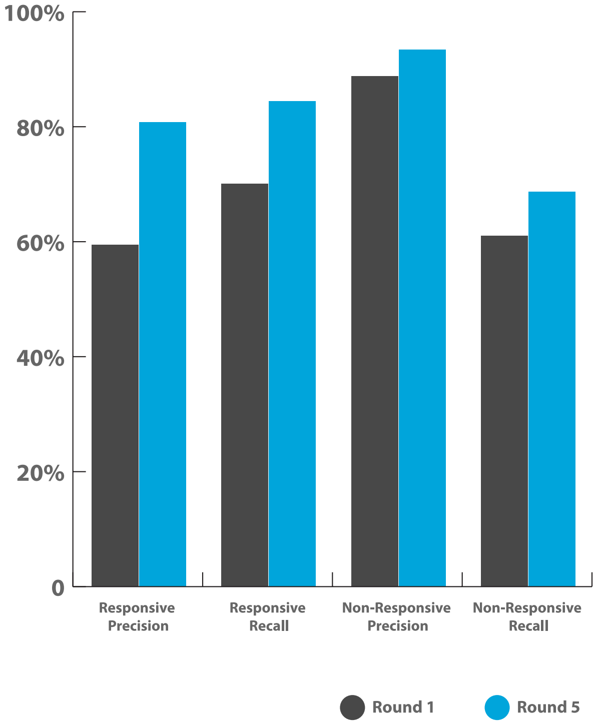

Figure 5: Change to Precision and Recall during Assisted Review Process

Conclusions

End users of a computer-assisted review workflow need to have confidence in two aspects of the Assisted Review workflow: first, that Assisted Review reporting using sampling accurately reflects the full population, and second, that document categorization improves with each round. The reasonableness of a document review may be determined by the reliance and transparency for these two aspects of the Assisted Review workflow.

With the current concern in the industry regarding statistics in computer-assisted review, we felt we needed to verify fairly straightforward statistics with questions one and two. For one, we needed to confirm that statistical sampling does indeed create representative samples. Given that we repeatedly fell within the margin of error, we can conclude that our samples were indeed representative.

Question two is a reformulation of question one, since sampling after categorization is no different than sampling before categorization. However, interpreting the results of summary reports is key for determining when to stop the workflow. Again, we found that sampling after categorization accurately reflected the proportion of documents correctly categorized.

As for kCura's overall Assisted Review workflow, we found that it improved effectiveness with almost each new round that was tried in our testing. It only took two or three responsive and non-responsive rounds to yield substantially improved results. Note that our results for question three are not as crystal clear as the results for question two, as effectiveness varies based on documents used, topics used, and the quality of the relevance judgments. However, we can certainly say for this particular dataset and these particular queries with these relevance judgments that the Assisted Review process improved effectiveness.

> **“As for kCura's overall Assisted Review workflow, we found that it improved effectiveness with almost each new round that was tried in our testing.”**

About the Authors

Dr. Grossman was an associate professor of computer science at IIT since 1999. He moved to England this past August and he is now the associate director of the Georgetown Information Retrieval Laboratory, a faculty affiliate at Georgetown University, and an adjunct professor at IIT. He has over 75 published papers and is the co-author of the book “Information Retrieval: Algorithms and Heuristics.” In 2013, he authored “Computer Science Programming Basics in Ruby,” which is available here. Dr. Grossman has taught undergraduate and graduate courses in information retrieval, data mining, and information security.

Acknowledgements

The author wishes to thank **Dr. Ophir Frieder**—Robert L. McDevitt, K.S.G., K.C.H.S and Catherine H. McDevitt L.C.H.S Chair in computer science and information processing at Georgetown University, and a professor of biostatistics, bioinformatics, and biomathematics at Georgetown University Medical Center—for his valuable assistance.

He’d also like to thank **Jay Leib** for providing guidance on Relativity Assisted Review throughout the process.

kCura®

231 South LaSalle Street, 8th Floor, Chicago, IL 60604
T: 312.263.1177 • F: 312.263.4351
sales@kcura.com • www.relativity.com