Further Discussion on FOMO

Herbert L. Roitblat, Ph.D.

Mimecast

In preparing this FOMO paper, comments were solicited from several experts, which were very helpful. After revising the paper, some additional comments were received that could not be incorporated into the paper, but do raise some interesting points for discussion.

The gist of the commentary is the idea that there could be facts that are missed using the approach that I outline. This argument is largely one over whether a glass is half full or half empty. Actually it is over whether the glass is 95% full, or 5% empty.

To review the process that I suggested, we find that at the end of our eDiscovery process, we have achieved 80% Recall. The approach does not care how that level of Recall was achieved or how it was measured. By whatever means, 80% Recall has been achieved. The approach makes note of the fact that 80% of the documents does not mean that 80% of the facts of the case have been discovered. Some facts are likely to appear in several documents, and some facts are likely to appear in only a few, but once a certain number of documents has been identified, the likelihood of finding additional facts among the documents that have not been identified as responsive will be low. There might be a disagreement about just how low that is and how low it should be, but the fact that it is low is apparently not in dispute.

The first approach I describe uses a 95% confidence level to determine a certain probability. This is the per document probability that a fact will be found with 95% probability. A 95% confidence level has its ordinary meaning that if we repeated the experiment 100 times, 95% of those would contain the parameter being estimated. This per document probability is the top of the confidence interval, whose bottom is 0.0.

Simulation verifies this prediction. If the identified set contains 10,000 responsive documents, for example, then the probability per document that would be found 95% of the time is 0.03%. That is, on average, we would find one fact with this probability for every 3,333 **responsive** documents that we read.

The comments pointed out that there could be facts that are more rare, say with a probability of 0.01%. We would need many more responsive documents in the identified set to achieve a confidence interval an upper bound of 0.01%. So, while other per document probabilities could be selected, their selection would be arbitrary rather than based on a 95% confidence level. After looking at, say, 10,000 documents in the identified set, we are justified at choosing 0.03%, but any other choice would be arbitrary.

A fact with a per-document probability of 0.01% would be found on average once per 10,000 **responsive** documents. When you recognize that these per document probabilities are per **responsive** documents, which are only a small percentage of all of the documents in the collection, the effort required to find at least one additional fact outside of the identified set balloons. In my opinion, it would be an extraordinary case that would merit such an effort.

Given the number of **responsive** documents in the missed set, there is a reasonable probability of finding a fact with that probability in the missed set. But in order for that fact to be a new one, it has to have been absent from the identified set and found in the missed set. So, if we knew which documents were responsive and outside of the identified set we would have a reasonable probability of finding a fact with that probability. But, for the fact to be new, it has to **not** be found in the identified set and it has to **be** found in the missed set, and there is a low probability of this conjunction, made even lower because in actual use, we do not know which are the responsive documents outside of the identified set.

There are a couple of other issues that the commentators raised that I want to consider. One of them revolves around just what I mean by a fact. I have in mind the kind of idea that is called a topic in the information retrieval field. The notion is that documents are generated from some set of topics or facts. Each fact is associated with a distribution of words. So a document is generated by first selecting a topic with a certain probability and then selecting a number of words from this topic each with a probability that depends on the topic. There can be a lot of overlap in the words associated with each topic, so the words are at best imperfect indicators of the topics that generated them. Although the words can vary from document to document in highly nuanced ways, the same words can express different topics and same topics can be associated with different words.

The commentators raised the question of whether the number of facts that I have chosen for my example might be too low. Part of this disagreement seems to stem from their understanding (I would say misunderstanding) of just what a fact is in this framework.

In my view, the facts differ in their probability of occurrence and in their importance. These are not the same thing. I have heard many times from a number of lawyers what may a canard, that they can prove any case with under a dozen documents. In my years of working with eDiscovery review, the number of topic tags that have been set up is usually limited to a couple of dozen. One case set up 300 tags, but the reviewers only used a few of these reliably. Finally, Latent Semantic Analysis (a technology that now belongs to Relativity) often used 300 dimensions in its semantic analysis. None of this proves that 300 facts is enough, but it does suggest that 300 is a reasonable number. The number of bins (representing topics) and the exact probabilities associated with each one can change without changing the overall pattern of the conclusions.

The final point of their critique that I want to raise concerns the potential for non-random selection of documents. The FOMO analysis assumes for simplicity's sake that facts are randomly assigned to documents and randomly assigned to the identified and missed sets. A substantial departure from randomness could significantly affect the pattern of conclusions. I think that this potential bias is a worthy topic of investigation, but without further evidence, I do not have any reason to think that it is a significant factor in this analysis.

None of the commonly used methods for distinguishing between responsive and non-responsive documents has access to or even tries to estimate the underlying topics or facts as the FOMO analysis uses that concept. The various systems may be non-random with

respect to the words of the documents, but we do not know how effective they are at selecting the facts of the documents. I have no reason to argue that continuous active learning, for example, is ineffective at identifying responsive documents, but I don't believe that it is effective at selecting facts. It is not designed to identify facts and we have no information at all about whether it nonrandomly assigns some facts to the identified set and leaves others to the missed set. Finally, we do not know how much nonrandom bias is too much bias for the FOMO framework. All models involve simplifying assumptions and it may turn out that the most used methods of document selection do affect the probabilities of topic selection, but that will have to be a subject for further investigation.

It is easy to get bamboozled by mathematical arguments, whether I make them or someone else does. But the basic ideas of the FOMO analysis are reasonably straightforward. Facts are not the same thing as documents but can occur in multiple documents. The more responsive documents you have, the more likely you are to have more of the facts. The cost of looking for additional facts goes up and the probability of finding more of them goes down with more documents. Armed with this information, one can differ on exactly where the cutoffs should be, but at some point, the value obtained from additional searching falls below the cost of doing that additional searching. The FOMO analysis provides a framework for understanding that tradeoff.