FOMO and eDiscovery

Herbert L. Roitblat, Ph.D.
Mimecast

"**Fear of missing out**, or **FOMO**, is 'a pervasive apprehension that others might be having rewarding experiences from which one is absent'." –Wikipedia

One of the successes from the introduction of machine learning to eDiscovery was the recognition that the accuracy of the document identification process could be measured. In fact, it has always been possible to measure the correctness of any document classification system (for example, responsive versus nonresponsive), even one conducted solely with human review, but that fact became easier to see in the context of justifying the use machine learning.

But the availability of eDiscovery measurement also contributed to conflict and anxiety. Legal professionals always knew that their eDiscovery processes were less than perfect, but now they could measure just how imperfect they were. Some parties were unable to ignore the imperfection, using it either as a weapon to force more work on their opponents or as a source of worry. If a producing party reported that their eDiscovery process found 80% of the responsive documents (80% Recall), is it not reasonable to expect them to go look for the remaining 20%? Might there not be some critical information among the documents that were missed? The short answer, is no. It is unlikely that such an effort will yield significant new information that was not already apparent in the 80% of the responsive documents that were correctly identified. Understanding why is actually pretty easy.

To explain how this process works, let us construct a situation in which we know exactly which documents were truly responsive and which were not. This imaginary world would allow us to identify the statistical properties of collections and would be free of any inaccuracies due to human foibles and errors, due to lack of agreement on which documents are responsive, and free of any limitations of our machine learning systems to identify responsive documents.

Further, let's assume that the goal of eDiscovery is to identify the facts of the case. Each document consists of a set of facts. Many documents may contain the same facts, and after identifying one of them, there is diminishing value in identifying others. For now, let's simplify the situation further by assuming that each document contains only one fact. Multiple facts per document just makes the process of finding these facts even easier than we will see that it is.

Responsive documents are those that contain responsive facts. Some facts inevitably occur more often than others. Some of them are very likely, and some are rare. In what follows, we are only concerned with the responsive documents—the 80% that are assumed to have been identified and the 20% that were predicted to have been missed.

With this statistical background, we can now ask whether there is a significant risk that there will be new facts in the documents that were missed (the missed set), facts that were

not in the documents identified as responsive (the identified set). By definition, any project with less than 100% Recall has missed certain documents, but, it turns out, it is unlikely that it has missed facts.

There are at least two ways to think about this problem. One way assumes that the number of facts that make a document responsive is unlimited. The other way assumes that only a certain number of facts is relevant to making a document responsive or valuable. In both cases, the analysis concerns only the truly responsive documents, which, obviously, are a subset of the total number of documents in the collection. I will try to describe these approaches more or less intuitively, leaving the mathematical details to an appendix.

Approach 1, the probability of a novel fact in the missed set

If we looked only at the identified set of responsive documents, then any facts that are missing from this set would never be available to the case. On the other hand, for the missed set to contain a novel fact, that fact would have to be absent from the identified set and present in the missed set. We can compute the probability that a fact is missing from the identified set and present in the missed set.

Not surprisingly, the more documents there are in the identified set, the more likely we are to find a rare fact among them. We can compute just how unlikely a fact would have to be to be absent in the identified set. Similarly, we can compute the probability that such an unlikely fact would be found in the missed set. When you combine the probability of missing a fact in the identified set and the probability of finding that fact in the missing set, the odds of finding a **new** fact in the missed set are low.

At 80% Recall, with a 95% confidence level, there is 5% chance that a fact will be absent from the identified set (100% - 5%). Conversely, there is a 52.7% chance a fact as rare as predicted will be found in the missed set. So overall, there is approximately a 2.6% chance that a fact will both be absent from the identified set and found in the missed set. Said another way, there is a 97.4% probability that no novel facts will be found in the missed set. These percentages apply no matter how many responsive documents there are, assuming that we have achieved 80% Recall, because of the ratio of the sample sizes in the two sets. Similar relationships apply at other levels of Recall.

Note that the prediction of a 2.6% chance of finding a new fact is not 2.6% of documents that will have a missing fact, but 2.6% of collections. On average, only 1 in 38 collections will have one or more novel facts among the missed set documents.

The odds are strongly against finding any new facts by examining the responsive documents that were not already identified. Of course, in reality, the responsive documents are only a small percentage of the documents in the collection. If we do not have the advantage of already knowing which documents are responsive, then we would have to read a lot of non-responsive ones to find these rare responsive facts. Eighty percent Recall may tell us that we have missed 20% of the responsive documents, but it does not tell us which ones they

are, so all of the remaining documents would have to be analyzed using one method or another.

Approach 2, Covering the consequential facts

The preceding analysis assumes that there is an unlimited number of potential facts that could be discovered. No matter how big the set of identified responsive documents is, there could still be a fact left to discover. That fact may be very rare, but it could still be out there. If we are willing to assume that there may be a fixed number of potential facts, then once those facts have been discovered in the identified set, there would be no new ones left to find in the missed set. We can then ask the question of how many responsive documents do we have to identify to find all of the facts?

How many facts are likely to be consequential in a case? A dozen, a hundred? Once these facts have been identified, is there really any reason to look at additional documents? Let's assume, somewhat arbitrarily, that there are 300 consequential facts.

Let's further assume that these facts differ in probability, some are in more documents than others and the frequency of facts follows the same kind of pattern as the frequency of words, called Zipf's law. Some facts are repeated many times and most of them occur with a much lower frequency. Some may be very rare.

With these two assumptions, we can predict the number of responsive documents that would have to be identified in order for all of these facts to be identified with high probability. In fact, the number of documents that would be needed to find all 300 consequential facts is surprisingly small. In the best case, it requires 2,394 documents. In the worst case, it requires 17,097 with the current set of assumptions. The range of these estimates is quite broad. Moreover, they are highly dependent on the specific assumptions used to calculate them, which are in the appendix. Nevertheless, this kind of result does show that with reasonable assumptions, the number of documents needed to find a reasonable set of facts is actually rather modest and within the range of many existing eDiscovery processes. This analysis depends on the number of documents in the identified set, but does not depend on the level of Recall.

Discussion

Statistical claims depend strongly on the assumptions by which the statistics were generated. Both approaches described here rest on the identified responsive documents being chosen randomly from the set of all responsive documents. Is that assumption reasonable?

The concept of a randomly selected set of documents means that every document has an equal chance of being the one containing the fact in question. From these analyses we still do not know just what fact that is. It could be anything. But to select documents non-randomly with respect to that fact, we would have to know just what it is and find a scheme that would allocate it to the missed set. In the face of ignorance about the identity of a missed fact, there may not be a plausible way that the two sets could be constructed that

would violate the randomness assumption. These results, therefore, would appear to apply to any good-faith methods used to identify the responsive documents.

On their face, not every eDiscovery process selects documents randomly so there is no guarantee that each fact has an equal chance of being selected. That does not mean, however, that this analysis is wrong for these approaches. We do not know how much these approaches bias the selection of facts in reality or what the practical consequences of that violation are. These methods operate by selecting words, but each word could be associated with multiple facts.

Many statistical and machine learning approaches are robust relative many of their assumptions. For example, one approach, Naïve Bayesian Classifiers assumes that words are distributed independently of one another. But, it is obvious that words are correlated. If the word "judge" appears in a document, then words like "lawyer" are more likely to also appear than words like "elephant." Nevertheless, Naïve Bayesian Classifiers are often said to be "surprisingly effective."

At this point, we don't know just how much the common ways of selecting documents for inclusion in the Recall set affect the predictions described here. However, I would not expect these approaches to be sufficiently selective relative to the underlying facts, as we use the concept, to affect strongly the results described here.

I made the simplifying assumption that each document contained only one fact, but we know that documents can contain more than one fact. If documents contained an average two independent facts, that would further raise the probability that a fact would be found in the identified collection, making it more unlikely that a surprising fact would be found only in the missed set. Because the first approach does not depend on the number of documents, the probabilities would remain the same if the number of facts per document increased.

For the second approach, fewer documents would have to be found to identify all of the relevant facts because each document could provide either one or two facts.

This pattern of results from both approaches suggests that it may not be necessary or reasonable to push further to inspect documents that may have been missed through the discovery process. Documents may have been missed, but it is unlikely that a significant number of facts have been. As a result, the remaining documents are very likely to be duplicative of those already identified.

The challenge of finding additional relevant facts in the missed set is likely to be further exacerbated by the difficulty of identifying relevant documents. The producing party would have to search not just the remaining responsive documents to find additional facts, but all of the so-far uncategorized documents. Many non-responsive documents would have to be considered in order to identify the few potential new facts that might be included in the missed set.

It is important to keep the numbers mentioned in this paper in perspective. As Robert Abelson explains, statistics should be viewed as principled story telling. The numbers here depend on the assumptions made, including the simplifying assumptions. They would certainly vary when we add in the variability of human judgments about the documents. These numbers should not be taken too literally, but they do tell a principled story—that finding 80% of the responsive documents does not mean that you have found only 80% of the facts and there is likely to be little value, but a lot of burden, in searching the remainder.

Math appendix

Understanding the details of the math is not necessary to understand the arguments of this paper, but some readers may be interested in specifically how these claims were constructed.

This example is built around a Recall level of 80%. Like many of the assumptions in this paper, these numbers are used illustratively, not prescriptively. Any actual case is likely to differ.

At 80% Recall, if there are 10,000 documents in the identified set, there are 2,500 documents in the missed set (10,000 / (10,000 + 2,500) = 0.8). At a 95% confidence level, we can say that a fact that was not seen in 10,000 responsive documents has a probability (prevalence) of 0.0003 or less per responsive document (eq. 1). That means that on average, we would have to read over 3,000 *responsive* documents to find even one containing this fact. The missing set in this example contains 2,500 documents, and the chance of finding the fact there is about 52.7%. So, for a fact to be missed in the identified set and found in the missing set, the fact would have to be very rare. At 95% confidence, this fact would be absent from the identified set in about 5% of the collections, and it would be identified in the missing set in about 52.7% of collections. Therefore, we would find a new fact in the missed set in only about 2.6% of all collections (5% x 52.7%).

Confidence level is the probability that the parameter estimated from the sample is contained within the confidence interval. Because we are interested in the probability of observing 0 occurrences of some event in the identified set, the confidence interval must include 0, with an upper bound at the specified prevalence. After n observations, we have the probability of not finding the occurrence in the first observation, not finding it in the second, and so on or $(1\text{-}p)^n$. The probability that the item we are looking for is not in the collection at some point reaches 1 - c. That is, it is exactly the probability of the event in the set, given that we have not observed it after looking for it *n* times. The parameter *p* is the probability of observing the event per observation, c is the probability of observing the event per collection. It is the probability that the actual prevalence is within the confidence interval bounded by 0.0.

The lower the level of Recall, the more likely we are to discover a new fact in the missed set, but this increase is rather small. At 70% Recall, the probability of finding a new fact increases to 3.6%. At 60% Recall, it increases to 4.3%.

The first two equations concern the first approach with potentially unlimited facts.

In the first equation, *p* is the maximum probability of a fact that has not been found in the Identified set. c is the confidence level used (95%). Ln is the natural log function, and n_1 is the number of documents in the identified set, n_2 is the number of documents in the missed set.

1

$$ln(1-p) = \frac{ln(1-c)}{n_1}$$

In the next equation, *q* is the probability of finding a fact with probability *p* from Eq. 1, and n_2 documents, the number of documents in the missed set.

2

$$q = 1 - (1-p)^{n_2}$$

Eq. 3. concerns the second approach. It describes how to estimate the expected number of recognized facts, given that a fixed number of documents has been reviewed. Statisticians talk about this as a bins and balls problem. By analogy, the documents in our identified set are "balls" that we throw at the facts, the "bins." We throw one ball at a time, and that ball falls into bin_i with probability p_i. The probability of each fact/bin is distributed according to the power law with an exponent, β, = 1.07: The choice of β = 1.07 is somewhat arbitrary, but it must be greater than 1.0. This value has been used in analyses of other phenomena and is reasonable, but not definitive.

3

$$p_i = \frac{1.0}{i^{\beta}}$$

The expected number of balls required to fill *n* bins with at least one ball is difficult to calculate. Instead we can describe the expected lower limit of this number and the expected upper limit.

$$E(T) \geq \frac{\left(n - \sqrt{n}\right)^{\beta}}{c_n^{\beta}} ln\left(\sqrt{n} + 1\right)$$

4

$$E(T) \leq \frac{n^{\beta}}{c_n^{\beta}} \left(1 + ln(n)\right)$$

Where

4

$$c_n^{\beta} = \frac{\beta - 1}{\beta - n^{1-\beta}}$$

Is the normalizing factor.

References

Robert P. Abelson. 1995. Statistics as principled argument, Lawrence Erlbaum Associates.

Ioannis Atsonios, Olivier Beaumont, Nicolas Hanusse, Yusik Kim. 2011. On Power-Law Distributed Balls in Bins and its Applications to View Size Estimation. ISAAC, Dec 2011, Yokohama, Japan. 2011. https://hal.inria.fr/inria-00618785/document

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. J. ACM 57, 2, Article 7 (February 2010), 30 pages. https://arxiv.org/abs/0710.0845

R. M. Cannon, & R. T. Roe, 1982. Livestock Disease Surveys – A Field Manual for Veterinarians. Canberra.

Jette Christensena, & Ian A. Gardnerb. 2000. Herd-level interpretation of test results for epidemiologic studies of animal diseases. Preventive Veterinary Medicine 45 (2000) 83-106. ftp://s173-183-201-52.ab.hsia.telus.net/AgroMediaDocs/JDrefs/PVM45_83.pdf

Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming and F. J. Smith 2003. Extension of Zipf's Law to Word and Character N-grams for English and Chinese Computational Linguistics and Chinese Language Processing Vol. 8, No. 1, February 2003, pp. 77-102. http://www.aclweb.org/anthology/O03-4004

Herbert L. Roitblat. 2007. Search and information retrieval. The Sedona Conference Journal. 8 (2007) Fall, 225-238.

Herbert L. Roitblat, Anne Kershaw, and Patrick Oot. 2010. Document categorization in legal electronic discovery: computer classification vs. manual review. J. Am. Soc. Inf. Sci. Technol. 61, 1 (January 2010), 70-80.

Issei Sato and Hiroshi Nakagawa. 2010. Topic models with power-law using Pitman-Yor process. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10). ACM, New York, NY, USA, 673-682. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.592.4256&rep=rep1&type=pdf