Sampling Strategies and Active Learning for Volume Estimation

Haotian Zhang, Jimmy Lin, Gordon V. Cormack, and Mark D. Smucker
University of Waterloo, Ontario, Canada
{haotian.zhang, jimmylin, gvcormac, mark.smucker}@uwaterloo.ca

ABSTRACT

This paper tackles the challenge of accurately and efficiently estimating the number of relevant documents in a collection for a particular topic. One real-world application is estimating the volume of social media posts (e.g., tweets) pertaining to a topic, which is fundamental to tracking the popularity of politicians and brands, the potential sales of a product, etc. Our insight is to leverage active learning techniques to find all the "easy" documents, and then to use sampling techniques to infer the number of relevant documents in the residual collection. We propose a simple yet effective technique for determining this "switchover" point, which intuitively can be understood as the "knee" in an effort vs. recall gain curve, as well as alternative sampling strategies beyond the knee. We show on several TREC datasets and a collection of tweets that our best technique yields more accurate estimates (with the same effort) than several alternatives.

1. INTRODUCTION

Suppose we would like to estimate the number of relevant documents in a collection for a particular topic. We refer to this as the volume estimation problem. How would we go about doing this both *accurately*, such that our estimate is as close as possible to the actual value, and *efficiently*, with as little effort as possible? This problem presents an interesting twist on the problem of high-recall retrieval. For example, in electronic discovery [6], the litigants are interested in the *actual* documents, whereas we just want the *volume* of the relevant documents. Of course, if we can find all the relevant documents, we can just count them—so existing active learning techniques for high-recall retrieval provide a baseline. Alternatively, we could just randomly sample from the collection to estimate the prevalence and infer the volume. The question is: can we do better than either approach?

Even assuming we can, who cares? Under what circumstances would we like to know the number of relevant documents without identifying the actual relevant documents? The concrete instantiation of our problem is estimating the volume of social media posts (e.g., tweets) pertaining to a topic. Volume estimation in this case is fundamental to several real-world applications: tracking the popularity of politicians and brands, the box-office appeal of movies, the audience for a sporting event, the potential sales of a product, and so on. The number of tweets pertaining to unfolding events is an often-quoted statistic in media coverage. Thus, volume estimation is not only academically interesting, but has significant real-world value.

The contribution of this paper is the development and evaluation of a technique for volume estimation based on active learning and sampling. Consider an active learning approach that can be characterized by a gain curve. The "knee" of that curve corresponds to the point where all the "easy" documents have been found. Our idea is to take advantage of active learning until the knee, and then use sampling techniques to extrapolate on the remainder of the collection. We present a simple technique for finding the knee that works well in practice and explore three different sampling approaches past the knee. On several TREC datasets and a collection of tweets, we show that our stratified sampling technique yields the most accurate estimates compared to other techniques with the same amount of effort.

2. BACKGROUND AND RELATED WORK

The volume estimation problem is related to retrieval techniques that focus on achieving very high recall—motivating application domains include legal eDiscovery, systematic reviews in evidence-based medicine, and locating prior art in patent search. The starting point of our work is the so-called baseline model implementation (BMI) that was provided to participants of the TREC 2015 Total Recall track, whose principal purpose was to evaluate, through a controlled simulation, methods to achieve very high recall—as close as practicable to 100%—with a human assessor in the loop. BMI is based on the "AutoTAR" technique of Cormack and Grossman [3]. Starting with a query, AutoTAR applies continuous active learning using relevance feedback to prioritize documents for human assessment. As in standard active learning, relevance judgments incrementally refine an underlying relevance model—but unlike the standard formulation of active learning, the point of AutoTAR is not to arrive at the best decision boundary between relevant and non-relevant documents, but rather to find all the relevant documents within a finite collection.

The effectiveness of techniques for high-recall retrieval is typically characterized by gain curves plotting effort (x-axis) vs. recall (y-axis). Most techniques, including the BMI and

SIGIR '16, July 17 - 21, 2016, Pisa, Italy

ISBN 978-1-4503-4069-4/16/07…$15.00
DOI: http://dx.doi.org/10.1145/2911451.2914685

even manual approaches, exhibit gain curves that begin with a "ramp up" phase, followed by a period where gain rises steadily, indicating the continuous discovery of "easy to find" relevant documents, and end with a region where the gain curves level off as the remaining relevant documents become more difficult to find. The "knee" is where the gain curve levels off, and this is a feature we exploit in our solution. In this work, we use the BMI from the TREC 2015 Total Recall track as-is, with the enhancement of a stopping criterion.

3. APPROACH

Suppose we would like to estimate the number of relevant documents in a particular collection consisting of D documents. How might we go about doing this?

A naïve approach might be to randomly sample (without replacement) documents from the collection and assess them for relevance. We can approximate this as a Bernoulli process, for which the volume estimate R_T is $D \cdot R_E/E$, where we find R_E relevant documents after examining a total of E documents. Note that this does not form a complete algorithm because we still need to know how many samples to draw. We return to this issue later when explaining our paired experimental methodology.

An alternative starting point might be to leverage an existing high-recall retrieval technique such as the BMI (Section 2). That is, we simply *find* all the relevant documents, which is a trivial way to determine the count. The problem, however, is that the BMI also does not come with a stopping criterion. Although various researchers have tackled this and related issues [7, 1], it is by no means solved.

Regardless, let's say we apply BMI to judge A documents. During this process, we will have explored some fraction of the collection that is more likely to contain relevant documents, and say we discover R_A documents. This value, of course, provides a lower bound on the number of relevant documents in the collection. But the problem is that we don't know how many relevant documents there are in the documents we didn't look at (what we refer to as the residual collection). Nevertheless, we can estimate an upper bound using the rate at which we're finding relevant documents just before we stopped, and extrapolate to the remainder of the collection. However, this makes a terrible assumption because any active learning technique will prioritize documents more likely to be relevant before documents less likely to be relevant, and so such an extrapolation would yield an unrealistically large overestimate.

One solution is to sample the residual collection. This requires answering two questions: First, how do we find the knee (i.e., the value of A)? Second, how do we sample after the knee? We tackle these two questions in turn.

3.1 Find the Knee

In our approach, we employ the BMI augmented by the following the knee-finding method proposed by Cormack and Grossman [4]. In each iteration, the BMI selects exponentially larger batches of documents for human judgment. After receiving feedback for each batch, we can trace the gain curve described in Section 2; to be precise, the y axis is now the number of relevant documents found and not recall.

At the end of each iteration, we have a point that corresponds to the total number of relevant documents found and the total number of documents judged thus far. We propose a candidate knee point as follows: find a point on the gain curve with maximum perpendicular distance from a line between the origin and the current (i.e., last) point of the curve. Let p_0 be the slope of the line from the origin to the candidate knee, p_1 be the slope of the line from the candidate knee to the last point, and the slope ratio $\rho = \frac{p_0}{p_1}$. We accept the candidate knee (and terminate) if: (1) the number of documents examined exceeds 1000 (to ensure that the active learning process has gotten beyond the "ramp up" phase), and (2) $\rho > 6$, if at least 150 relevant documents have been retrieved, or $\rho > 156 - r$, if $r < 150$ relevant documents have been retrieved. The second clause in (2) is a special case for handling topics with few relevant documents. The parameters for this technique were tuned on a private dataset.

Note that to be precise, this technique doesn't actually discover the knee until we've "passed" the knee, but the intuition nevertheless holds. The actual switchover point where we stop active learning and begin sampling corresponds to the point where we discovered the knee. However, for expository convenience we still refer to "stopping at the knee".

3.2 Sampling Strategies

Based on the knee-finding algorithm described in the previous section, we apply the BMI until the stopping criterion is met. At that point, we have judged A documents and found R_A relevant documents. The next question is: what do we do with the residual collection that we have not yet examined? We present three sampling strategies:

Negative Binomial Sampling. In this approach, we sample from the residual collection until we encounter M relevant documents; let's say this process requires us to examine S documents. Each sample can be modeled as a Bernoulli trial, and thus the sampling process can be characterized by a negative binomial distribution. Under this interpretation, the minimum variance unbiased estimator for $\hat{p}$, the probability of success (i.e., probability of a document being relevant) is given as $\hat{p} = (r-1)/(r+k-1)$, where r is the number of successes (relevant documents, which we set to M) and k is the number of failures (non-relevant documents) in our sequence of observations [5].

From this, our estimate of the total number of relevant documents, R_T, is as follows (note $S = r + k$):

$$R_T = R_A + (D - A)\frac{(M-1)}{(S-1)}, \text{ for } M > 1. \tag{1}$$

In our experiments, we tried setting $M \in \{2, 4, 8\}$. Naturally, higher values of M reduce the variance, but at the cost of requiring more assessment effort. The total effort required with this approach is $A + S$, where S is the number of documents we must assess to find M relevant documents.

The Horvitz-Thompson Estimator. The downside of the negative binomial sampling strategy is that for cases where the prevalence of relevant documents in the residual collection is low, it might require a lot of effort to find M relevant documents. An alternative is to use the BMI to score all documents in the residual collection at the point when it terminates, thus ordering all remaining documents in decreasing probability of relevance.

From here, we can apply a standard sampling technique called the Horvitz-Thompson Estimator (HT estimator) [8]: we compute a distribution over all documents in the residual collection such that its probability of being sampled is proportional to its probability of relevance (as estimated by the BMI). This renormalized distribution is referred to as the inclusion probability, i.e., π_i refers to the probability that document i will be sampled. The Horvitz-Thompson estimate of the number of relevant documents is:

$$R_T = R_A + \sum_{i=1}^{n} \pi_i^{-1} Y_i \tag{2}$$

where Y_i is an indicator variable for relevance in each of the n sampled documents. Note that this does not form a complete algorithm because we are missing a stopping criterion. Once again, we address this issue in our paired experimental methodology below.

Stratified Sampling. We propose a novel stratified sampling strategy, also based on a relevance ranking of the residual collection at the point when the BMI terminates. This approach proceeds in iterations: at the i-th iteration, we randomly sample K^S documents ($= 1000$) from the next top ranking K documents ($= 10000$) and judge those documents. Suppose we find R_i relevant documents: we can then estimate that there are $K \cdot (R_i/K^S)$ in the top K hits. We then proceed to the next iteration and sample K^S documents from the *next* K top ranking documents, repeating as long as we find at least one relevant document. The total number of relevant documents can then be computed as:

$$R_T = R_A + \sum_{i=1}^{n} K \cdot (R_i/K^S) \tag{3}$$

where n is the number of iterations. The total effort expended is $A + n \cdot K^S$.

4. EXPERIMENTAL SETUP

To evaluate our volume estimation techniques, we used test collections from the TREC 2015 Total Recall Track [7]. In particular, we used three collections: the (redacted) Jeb Bush Emails (called "Athome1"), consisting of 290k emails from Jeb Bush's eight-year tenure as the governor of Florida (10 topics); the *Illicit Goods* dataset (called "Athome2") collected for the TREC 2015 Dynamic Domain Track, consisting of 465k documents from a web crawl (10 topics); and the *Local Politics* dataset (called "Athome3") collected for the TREC 2015 Dynamic Domain Track, consisting of 902k documents from various news sources (10 topics). The relevance assessment process is described in the track overview paper [7], but for the purposes of our study, it suffices to say that the evaluation methodology has been sufficiently validated for assessing the effectiveness of high-recall tasks (and thus these collections are suitable for our volume estimation problem). Finally, as a validation set, we evaluated our techniques on the Twitter collection described in Bommannavar et al. [2], who *exhaustively* annotated approximately 800k tweets from one day in August 2012 with respect to four topics: Apple (the technology company), Mars (the planet), Obama, and the Olympics. This dataset exactly matches our motivating application: how much "buzz" is there on social media about a particular topic?

Note that random sampling and the HT estimator approaches are not complete estimation algorithms since they lack a stopping criterion. Therefore, they are compared to negative binomial sampling and stratified sampling in a paired setup, where we evaluate how the techniques compare at the same level of effort. This models an A/B testing scenario where we have two parallel efforts proceeding at exactly the same pace assessing documents. When one technique terminates, we also stop the other. At that point, we ask: how do the two estimates compare?

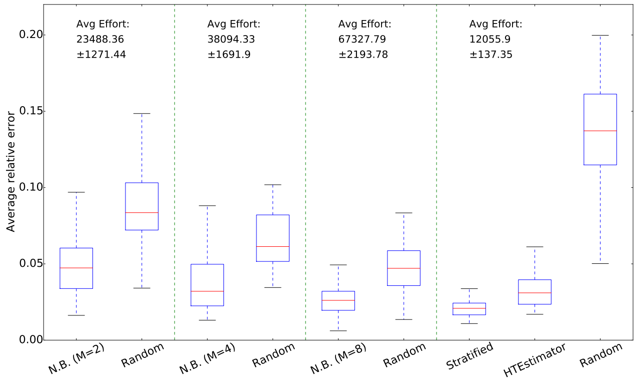

Figure 1: Box-and-whiskers plot characterizing 50 trials of each of our techniques on the Athome1 collection.

Our experimental procedure is as follows: for each topic in a collection, we ran our estimation technique (either negative binomial sampling or stratified sampling) and recorded the total effort. We then ran a paired experiment with either random sampling or the HT estimator (or both) using exactly the same level of effort. We recorded the estimated volume for all techniques. For each collection, we report the average (relative) error across all topics and the root mean square error. The Athome1 collection was used as our training set, on which we ran 50 trials of the above procedure to characterize the variability of estimates. The Athome2, Athome3, and Twitter collections were used as held-out test sets—we report the results of a single trial.

5. RESULTS

The results of 50 trials of our experimental procedure are shown in Figure 1, where the average relative error across the trials is characterized by a standard box-and-whiskers plot. We compared negative binomial sampling, $M = \{2, 4, 8\}$, with random sampling using the paired approach described above. We compared stratified sampling with the HT estimator and random sampling using exactly the same procedure. Each of these comparisons is shown by grouped bars (separated by dashed lines) in the figure.

As expected, the negative binomial sampling approach becomes more accurate with increasing values of M (but requires correspondingly more effort). For reference, the entire collection contains 290k documents, so with $M = 8$, on average we must examine nearly a quarter of the collection. However, we see that negative binomial sampling is more accurate than random sampling at the same level of effort. It is clear that our stratified sampling approach is superior to all other techniques. On average, stratified sampling requires about half as much effort as negative binomial sampling with $M = 2$ but gives much more accurate estimates. In fact, stratified sampling provides more accurate estimates than negative binomial sampling with $M = 8$, at about a fifth of the effort. Stratified sampling also beats both the HT estimator and random sampling at the same level of effort.

Results for Athome2 and Athome3 are shown in Table 1.

Measure	Avg Effort	Avg Relative Error	Root Mean Square Error
Athome2			
Neg. Binomial ($M = 2$)	80925	0.016	0.021
= sample	80925	0.094	0.122
Neg. Binomial ($M = 4$)	122527	0.014	0.023
= sample	122527	0.052	0.062
Neg. Binomial ($M = 8$)	181407	0.015	0.020
= sample	181407	0.045	0.060
Stratified	8363	0.026	0.042
= HTEstimator	8363	0.051	0.070
= sample	8363	0.410	0.621
Athome3			
Neg. Binomial ($M = 2$)	482237	0.041	0.105
= sample	482237	0.045	0.079
Neg. Binomial ($M = 4$)	546379	0.011	0.030
= sample	546379	0.042	0.073
Neg. Binomial ($M = 8$)	597489	0.023	0.058
= sample	597489	0.032	0.064
Stratified	3168	0.053	0.113
= HTEstimator	3168	0.100	0.200
= sample	3168	0.867	1.119
Twitter			
Neg. Binomial ($M = 2$)	24160	0.261	0.233
= sample	24160	0.222	0.240
Neg. Binomial ($M = 4$)	39162	0.106	0.090
= sample	39162	0.046	0.041
Neg. Binomial ($M = 8$)	40295	0.007	0.036
= sample	40295	0.179	0.181
Stratified	22687	0.047	0.048
= HTEstimator	22687	0.093	0.092
= sample	22687	0.170	0.218

Table 1: Results of various volume estimation techniques on the Athome2, Athome3, and Twitter collections.

Since these comprise our held-out test data, we only report the results of a single trial. In the table, rows are grouped in terms of different techniques at the same level of effort, e.g., the rows marked "= sample" denote accuracy with the same number of judged documents as the corresponding negative binomial or stratified condition. Due to the variability inherent in our sampling strategies, in our particular trial we observe greater error with $M = 8$ than with $M = 4$ using negative binomial sampling. This, however, is not inconsistent with the results in Figure 1.

Overall, the results on Athome2 (465k documents) are consistent with the results from Athome1, our training set. Negative binomial sampling becomes more accurate with increasing M and is more accurate than random sampling with the same level of effort. However, our stratified sampling technique provides comparable error at far less cost, beating both the HT estimator and random sampling.

Results on the Athome3 collection, which contains 902k documents, are quite poor. Table 2 shows why: for each topic in that collection, we list the total number of relevant documents, the effort expended in the active learning portion of our procedure, and the number of relevant documents found at that point. For five of the topics (those in bold), with active learning we've found either all or all but one of the relevant documents, which means that our termination condition for negative binomial sampling (e.g., with $M = 2$) is never met and hence the procedure forces us to examine *the entire collection*. In contrast, with stratified sampling we examine 1000 of the top 10000 documents, find zero relevant, and terminate.

The bottom of Table 1 shows our results for the Twitter collection. Once again, we report results from a single trial. Overall, the findings are consistent with the other collections: our stratified sampling technique clearly yields more accurate estimates than all other techniques. This gives us some degree of confidence that our algorithms, developed on email (Athome1), generalize to entirely different collections (tweets). We have no explanation as to why negative binomial sampling with $M = 4$ gives worse estimates than comparable random sampling, or why comparable random sampling with $M = 8$ gives such poor results. We purposely decided against error analysis in order to preserve the sanctity of this validation set.

Topic	Rel Docs	Knee Stop Effort	RelAtKnee
athome3089	**255**	1105	**254**
athome3133	**113**	1105	**112**
athome3226	2094	3478	2022
athome3290	**26**	2316	**26**
athome3357	629	1526	599
athome3378	**66**	1105	**66**
athome3423	76	1232	40
athome3431	1111	1232	1106
athome3481	2036	3478	1924
athome3484	**23**	1105	**23**

Table 2: Relevant documents identified and effort when BMI terminates for Athome3.

6. CONCLUSION

Estimating the number of relevant documents in a collection presents an interesting twist to the high-recall retrieval problem. Our results show that actually *finding* the relevant documents is a good approach to *counting* the total volume. However, we develop and verify the insight that we should first identify the "easy to find" documents and then extrapolate via sampling on the rest. Our approach establishes a baseline for future work on an important real-world application, particularly in the social media space.

Acknowledgments. This work was supported in part by the U.S. National Science Foundation under awards IIS-1218043 and CNS-1405688, the Natural Sciences and Engineering Research Council of Canada (NSERC), Google, and the University of Waterloo. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the sponsors.

7. REFERENCES

[1] M. Bagdouri, W. Webber, D. Lewis, and D. Oard. Towards minimizing the annotation cost of certified text classification. *CIKM*, 2013.

[2] P. Bommannavar, J. Lin, and A. Rajaraman. Estimating topical volume in social media streams. *SAC*, 2016.

[3] G. Cormack and M. Grossman. Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv:1504.06868v1*, 2015.

[4] G. Cormack and M. Grossman. Engineering quality and reliability in technology-assisted review. *SIGIR*, 2016.

[5] N. Johnson, A. Kemp, and S. Kotz. *Univariate Discrete Distributions, 3rd Edition*. Wiley, 2006.

[6] D. Oard and W. Webber. Information retrieval for e-discovery. *FnTIR*, 7(2–3):99–237, 2013.

[7] A. Roegiest, G. Cormack, M. Grossman, and C. Clarke. TREC 2015 Total Recall track overview. *TREC*, 2015.

[8] Y. Tillé. *Sampling Algorithms*. Springer Series in Statistics. Springer, 2006.