A Study on a Mixed Stopping Strategy for Total Recall Tasks

Giorgio Maria Di Nunzio[†,‡]
[†]Department of Information Engineering, [‡]Department of Mathematics
University of Padua, Italy
giorgiomaria.dinunzio@unipd.it

ABSTRACT

How do we calculate how many relevant documents are in a collection? In this abstract, we discuss our line of research about total recall systems such as interactive system for systematic reviews based on an active learning framework [4–6]. In particular, we will present 1) the problem in mathematical terms, and 2) the experiments of an interactive system that continuously monitors the costs of reviewing additional documents and suggests the user whether to continue or not in the search based on the available remaining resources. We will discuss the results of this system on the ongoing CLEF 2019 eHealth task.

CCS CONCEPTS

• **Information systems** → **Clustering and classification**; *Probabilistic retrieval models*; • **Applied computing** → **Health care information systems**; **Health informatics**.

KEYWORDS

Total recall; Probabilistic Models; Random Sampling

1 INTRODUCTION

Given a collection of documents and an information need, can we estimate the number of relevant documents in the collection for that information need? This question seems trivial but, for some tasks, a sufficiently accurate answer may help the user to save a lot of resources in terms of time and money. In fact, if we knew this number, we could stop the search process as soon as the last relevant document is found or we may decide to stop earlier if it is no longer convenient to continue the search.

The type of retrieval tasks that we refer to are, for example, e-Discovery [12] and Technology-Assisted Review (TAR) tasks [1] where one or more classifiers are trained using some manually annotated content in order to find the remaining relevant documents in the collection. Among many others, there are two key questions for these tasks: which documents should be chosen for manual review? When do we stop judging documents? The first question is usually addressed with an approach called Continuous Active Learning (CAL) [3] in which a retrieval system is continuously updated with the interactive feedback given by the user that is reading and judging the documents. The second question about the stopping strategy has been discussed in [2]. In the last years, international evaluation campaigns have organized experimental labs in order to evaluate systems designed to achieve very high recall through controlled simulation [7, 8].

In this abstract, we want to discuss our line of research that follows from the studies in interactive system for systematic reviews based on an active learning framework [4–6]. In particular, we present a system that continuously monitors the costs of reviewing documents and suggests the user whether to continue or not in the search based on the available remaining resources.

In order to avoid confusion with similar topics in the IR research field, we want to stress the fact that we are not studying whether the subset of relevant documents judged is sufficient to compare the accuracy of IR systems (like in TREC or CLEF) [11]; moreover, we are not proposing an alternative pooling strategy to build the set of relevance judgement [9] (although this approach may be extended in the future).

The paper is organized as follows: in Section 2, we present the problem in mathematical terms by means of the hypergeometric distribution to model the sampling of documents without replacement. In Section 3, we present a brief summary of the application of this approach to the ongoing CLEF 2019 eHealth task on Technology Assisted Medical Reviews.

2 MATHEMATICAL NOTATION

We assume to have a collection of N objects which can be classified as either relevant or non-relevant. The number of relevant objects is K; hence, the number of non-relevant documents is $N-K$. We draw n samples from the collection of objects without replacement and we want to compute the probability of observing k relevant objects. This probability function is represented by the hypergeometric distribution:

$$P(X = k; N, K, n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{1}$$

where X is the random variable with a hypergeometric distribution with parameters N, K, and n; $\binom{a}{b}$ represents the binomial coefficient.

For example, if we have a collection of $N = 100$ objects with $K = 20$ relevant objects, the probability of observing $k = 10$ relevant objects by drawing $n = 30$ objects from the collection is

$$P(X = 10; N = 100, K = 20, n = 30) = \frac{\binom{20}{10}\binom{80}{20}}{\binom{100}{30}} = 0.17 \tag{2}$$

If we could draw t times n samples from a hypergeometric distribution with parameters K and N, we would obtain a sampling

IIR 2019, September 16–18, 2019, Padova, Italy

distribution with mean $\overline{y}$ and standard error SE

$$\overline{y} = \frac{1}{t}\sum_{i=1}^{t}\frac{k_i}{n_i} \quad (3)$$

$$SE[\overline{y}] = \sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)} \quad (4)$$

where s^2 is the sample variance and $\left(1-\frac{n}{N}\right)$ the finite population correction factor ([10, Chapter 2]).

We can use the sampling distribution to compute how accurate our estimates of the mean are by means of confidence intervals which define the lower and upper bound of our estimate:

$$[\overline{y} - z_{\alpha/2}SE[\overline{y}], \overline{y} + z_{\alpha/2}SE[\overline{y}]] \quad (5)$$

where $z_{\alpha/2}$ is the (1 - $\alpha/2$)th percentile of the standard normal distribution. For example, if we drew $t = 10000$ times $n = 30$ objects from the hypergeometric distribution with $K = 20$ and $N = 100$, we would obtain a sampling distribution with mean $\overline{y} \simeq 0.1994$ and $SE[\overline{y}] \simeq 0.0094$. If we wanted a 95% confidence interval ($\alpha = 0.05$), the range of the number of relevant objects would be between 18 and 22.

3 EXPERIMENTS AND DISCUSSION

Our use case is building a system for systematic medical reviews which are a method to collect the findings from multiple medical studies in a reliable way. Given budget and time constraints, we need to provide the physician with a sufficient amount of (possibly all) relevant medical studies.

In such real life cases, we do not have a perfect knowledge about the collection of documents at our disposal: we may or may not know the exact number of documents N in the collection, or the exact number of relevant documents K, or both. In our experiments, we assume the following: 1) we know N, and 2) we know (for example an "oracle" tells us) that there are "at least" K^- relevant documents; we use a "minus" at superscript to indicate that this number is a lower bound for K. In other words, we know the total number of documents in the collection, but we have just a partial knowledge (a lower bound) on the number of relevant documents. Moreover, by "relevant" object we mean that some user has judged the document. Initially, we do not know whether K^- is close or not to the "true" value K; consequently, we want to estimate how costly it is to build a confidence interval for K^- accurate enough to tell whether to stop the search of additional relevant documents. In our previous example, suppose that the oracle says that there are at least $K^- = 20$ relevant documents in a collection of $N = 100$ objects. How many documents n do we need to draw (or read) to get a desired confidence interval?

The systems we propose uses a mixed approach to 1) find the relevant documents in a collections of medical documents given a query of a physician, and 2) compute the confidence interval of the estimate of the number of relevant documents left in the collection. On the one hand, we apply a Continuous Active Learning (CAL) approach using the BM25 ranking model [5]: i) the system ranks the documents in the collection and shows the top ranked document to the user; then, ii) the user reads the document and sends a feedback to the system (the document is relevant or not); finally, iii) the system is re-trained with this new piece of information and re-ranks the remaining documents. In this way, we aim to build the set of K^- relevant documents. On the other hand, every m ranked documents the system picks a random document from the collection and shows it to the user. In this way, we start building the confidence interval of the range of relevant documents that are still in the collection after having sampled n documents.

The system allows to adjust the proportion of documents that are sampled against those that are ranked in order to balance the costs of estimating the confidence intervals accurately versus finding the most relevant information as quick as possible. In particular:

- we set a number of documents d that the user is willing to read and a number m that tells the algorithm when to randomly sample a document from the collection;
- the first half of documents $d/2$ are used to continuously update the relevance weights of the terms according to the explicit relevance feedback given by the user;
- for the second half documents we use a Naïve Bayes classifier to select the subset of documents to read.

We will discuss the latest version of this system that participated in the last three editions of the eHealth CLEF 2019 lab and we will give some suggestions about future work that include query aspects [1] and the optimization of loss functions [12].

REFERENCES

[1] G. V. Cormack and M. R. Grossman. 2015. Multi-Faceted Recall of Continuous Active Learning for Technology-Assisted Review. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 763–766. https://doi.org/10.1145/2766462.2767771

[2] G. V. Cormack and M. R. Grossman. 2016. Engineering Quality and Reliability in Technology-Assisted Review. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, New York, NY, USA, 75–84. https://doi.org/10.1145/2911451.2911510

[3] G. V. Cormack and M. R. Grossman. 2016. Scalability of Continuous Active Learning for Reliable High-Recall Text Classification. In *Proc. of CIKM'16*. ACM, New York, NY, USA, 1039–1048. https://doi.org/10.1145/2983323.2983776

[4] G. M. Di Nunzio. 2018. Finding all the Needles in the Haystack. A System to Estimate the Costs of e-Discovery and Systematic Reviews. In *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems, Bertinoro, Italy, August 28-31, 2018.* 106. http://ceur-ws.org/Vol-2167/short9.pdf

[5] G. M. Di Nunzio. 2018. A Study of an Automatic Stopping Strategy for Technologically Assisted Medical Reviews. In *Proc. of ECIR 2018*. Springer, 672–677.

[6] G. M. Di Nunzio, M. Maistro, and F. Vezzani. 2018. A Gamified Approach to Naïve Bayes Classification: A Case Study for Newswires and Systematic Medical Reviews. In *Companion of the The Web Conference 2018 WWW 2018, Lyon , France, April 23-27, 2018.* 1139–1146. https://doi.org/10.1145/3184558.3191547

[7] M. R. Grossman, G. V. Cormack, and A. Roegiest. 2016. TREC 2016 Total Recall Track Overview. In *Proceedings of The Twenty-Fifth TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016.* http://trec.nist.gov/pubs/trec25/papers/Overview-TR.pdf

[8] E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker (Eds.). 2017. *CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview. In Working Notes of CLEF 2017. http://ceur-ws.org/Vol-1866/invited_paper_12.pdf*. Number 1866 in CEUR Workshop Proceedings. CEUR-WS.org.

[9] A. Lipani, M. Lupu, J. Palotti, G. Zuccon, and A. Hanbury. 2017. Fixed Budget Pooling Strategies Based on Fusion Methods. In *Proceedings of the Symposium on Applied Computing (SAC '17)*. ACM, New York, NY, USA, 919–924. https://doi.org/10.1145/3019612.3019692

[10] S.L. Lohr. 2019. *Sampling: Design and Analysis*. Taylor & Francis Group. https://books.google.it/books?id=8ezfwgEACAAJ

[11] Xiaolu Lu, Alistair Moffat, and J. Shane Culpepper. 2016. The effect of pooling and evaluation depth on IR metrics. *Information Retrieval Journal* 19, 4 (01 Aug 2016), 416–445. https://doi.org/10.1007/s10791-016-9282-6

[12] D. W. Oard, F. Sebastiani, and J. K. Vinjumur. 2018. Jointly Minimizing the Expected Costs of Review for Responsiveness and Privilege in E-Discovery. *ACM Trans. Inf. Syst.* 37, 1, Article 11 (Nov. 2018), 35 pages. https://doi.org/10.1145/3268928