

A Distributed Effort Approach for Systematic Reviews. IMS Unipd at CLEF 2019 eHealth Task 2.

Giorgio Maria Di Nunzio^{1,2}

¹ Department of Information Engineering

² Department of Mathematics

University of Padua, Italy

giorgiomaria.dinunzio@unipd.it

Abstract. This is the third participation of the Information Management Systems (IMS) group at CLEF eHealth Task of Technologically Assisted Reviews in Empirical Medicine. This task focuses on the problem of medical systematic reviews, a problem which requires a recall close (if not equal) to 100%. Semi-Automated approaches are essential to support these type of searches when the amount of data exceed the limits of users, i.e. in terms of attention or patience. We present a variation of the system we presented last year; in particular, not only we set the maximum amount of documents that the physician is willing to read, but we distribute the effort across the topics proportionally to the number of documents in the pool. We compare the results of this approach with the “frozen” system we used in 2018 and a BM25 baseline.

1 Introduction

In this paper, we describe the participation of the Information Management Systems (IMS) group at CLEF eHealth 2019 [2] Technology Assisted Review Task [1]. This task focuses on the problem of systematic reviews, that is the process of collecting articles that summarise all evidence (if possible) that has been published regarding a certain medical topic. This task requires long search sessions by experts in the field of medicine; for this reason, semi-automatic approaches are essential to support these type of searches when the amount of data exceed the limits of users, i.e. in terms of attention or patience.

The objective of our participation was to compare the system that we used in the previous year, with a new strategy to distribute the effort of the user (the physician or an expert in the field of medicine) across the topics. In particular,

- we re-use the stopping strategy to simulate the maximum amount of documents that a physician is willing to review in the two-dimensional approach presented in [5];

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2019, 9-12 September 2019, Lugano, Switzerland.

- we distribute the effort, in terms of number of documents to read, proportionally to the size of the pool of documents for each topic;
- we estimate the 95% confidence interval of the proportion of relevant documents present in the collection [6].

The source code of the experiments is available for reproducibility purposes.³

2 Approach

In this paper, we continue to investigate the interaction with the two dimensional interpretation of the BM25 model applied to the problem of explicit relevance feedback [9, 3, 8, 5, 7, 6].

In particular, the two-dimensional representation of probabilities [4, 9] is an intuitive way of presenting a two-class classification problem on a two-dimensional space. Given two classes, for example relevant \mathcal{R} and non-relevant $\mathcal{N}\mathcal{R}$, a document d is assigned to category \mathcal{R} if the following inequality holds:

$$\underbrace{P(d|\mathcal{N}\mathcal{R})}_y < m \underbrace{P(d|\mathcal{R})}_x + q \quad (1)$$

where $P(d|\mathcal{R})$ and $P(d|\mathcal{N}\mathcal{R})$ are the likelihoods of the object d given the two categories, while m and q are two parameters that can be optimized to compensate for either the unbalanced class issues or different misclassification costs.

We focused on the following problems:

1. study the effectiveness of a classifier given a fixed amount of documents that a physician is willing to review;
2. design a sampling strategy to estimate the 95% confidence interval of the number of relevant documents in the collection.

In the experiments, we used the same procedure we used 1st year [6]:

- we set a number n of documents that the physician is willing to read and a number s that tells the algorithm when (every s documents) to randomly sample a document from the collection instead of presenting to the physician the next most relevant document;
- for each topic, we run an optimized (hyper-parameters) BM25 retrieval model and we obtain the relevance feedback for the first abstract in the ranking list;
- from the second document until $n/2-1$, we continuously update the relevance weights of the terms according to the explicit relevance feedback given by the physician (simulated by the qrels available with the test collection);
- for the last half of the documents $n/2$ that the physician is willing to read, we use a Naïve Bayes classifier continuously updated with the explicit relevance feedback [5].

³ <https://github.com/gmdn/CLEF2019>

topic	pool	prop	shown
CD000996	281	0.003	43
CD001261	571	0.007	86
CD004414	336	0.004	51
CD006468	3874	0.047	583
CD007867	943	0.011	142
CD008874	2382	0.029	359
CD009044	3169	0.038	477
CD009069	1757	0.021	265
CD009642	1922	0.023	290
CD010038	8867	0.108	1335
CD010239	224	0.003	34
CD010558	2815	0.034	424
CD010753	2539	0.031	382
CD011140	289	0.004	44
CD011558	2168	0.026	327
CD011571	146	0.002	22
CD011686	9729	0.118	1464
CD011768	9160	0.111	1379
CD011787	4369	0.053	658
CD011977	195	0.002	30
CD012069	3479	0.042	524
CD012080	6643	0.081	1000
CD012164	61	0.001	10
CD012233	472	0.006	72
CD012342	2353	0.029	355
CD012455	1593	0.019	240
CD012551	591	0.007	89
CD012567	6735	0.082	1014
CD012661	3367	0.041	507
CD012669	1260	0.015	190
CD012768	131	0.002	20

Table 1: Proportion of documents per topic.

Instead of setting n equal for all topics, this year we tried a different approach in order to let the user to read more documents for those topics with more documents in the pool. In Table 1, we show, for each topic, the number of documents in the pool, the proportion of documents of the pool compared to the total number of documents pooled, the number of documents we will show to the user (to be multiplied by 2).

3 Experiments

For all the experiments, we set the values of the BM25 hyper-parameters in the same way we did in [6].

3.1 Official Runs

We submitted runs for three different systems:

- a BM25 baseline with continuous active learning and a fixed threshold for each topic,
- the “frozen” system fo 2018 with different proportions of documents to be read for the initial phase but with a fixed threshold for each topic,
- the new approach with a different threshold for each topic.

In particular, for the frozen system, we used 10% or 50% of the initial pool of documents per topic to build the classifier. The new distributed effort approach uses 10% of the pool at the beginning of the training, but, in general, it may stop earlier compared to the other approach if the effort required for a topic is low in terms of documents allowed.

3.2 Unofficial Runs

In order to compare the BM25 model with a similar proportion of documents shown to the user, we added some BM25 runs and removed some others that showed a different number of documents.

3.3 Evaluation Measures

In order to evaluate the performance of the systems, we chose the number of documents shown to the user as one of the performance measures since, in our case, it is also the point where we stop retrieving documents. In addition, we use recall and averaged recall across topics to measure the accuracy of the retrieval.

3.4 Results

In Figures 1 and 2, we show a topic by topic comparison of groups of runs: BM25, distributed effort, orginal 2018 with 10% or 50% of the initial pool selected. By increasing the threshold of the number of documents shown to the user, we are able to tune the performance of the system and reach an average recall close to 100% for all the systems under evaluation. Some topics are much more difficult than others; for example, topic CD011558 requires the retrieval of most of the pooled documents in order to achieve a reasonable recall (around 0.8).

In Figure 3, we show the performance of the four groups of runs in terms of average recall (across topics) given the number of documents shown to the user. By increasing the number of documents (from left to right) the four approaches increase the average recall and go beyond 90% even with less than 4% of the total number of documents, for example the two 2018 approaches of the frozen system.

The distributed effort approach we proposed this year performed worse than expected. It seems that by reducing the number of documents allowed per topic

too much, especially for topics with smaller pools, we obtain a suboptimal system compared to the original one. In other terms, it may be more convenient to set up a fixed cost per topic and use all the documents of the pool if necessary, instead of saving some resources for topics with more documents in the pool.

4 Conclusions

In this work, we presented a variation of the continuous active learning approach used in [6] that uses a fixed stopping strategy to simulate the maximum amount of documents that a physician is willing to review and a sampling strategy that is used to estimate the number of relevant documents in the collection. The result of the distributed effort approach were worse than expected, compared to the original approach in presented in 2018. The performance of the new system is still remarkable since it achieves an average recall of 90% by using only 10% of the documents in the collection; however, the original system can achieve the same results by reducing the number of documents shown to the user by half.

We are currently analyzing the results provided by the organizers and adding to the official runs a set of unofficial runs that will complete the picture of all the possible settings. As future work, we will study a methodology to dynamically vary the amount of documents according to the estimate of the amount of relevant documents still missing.

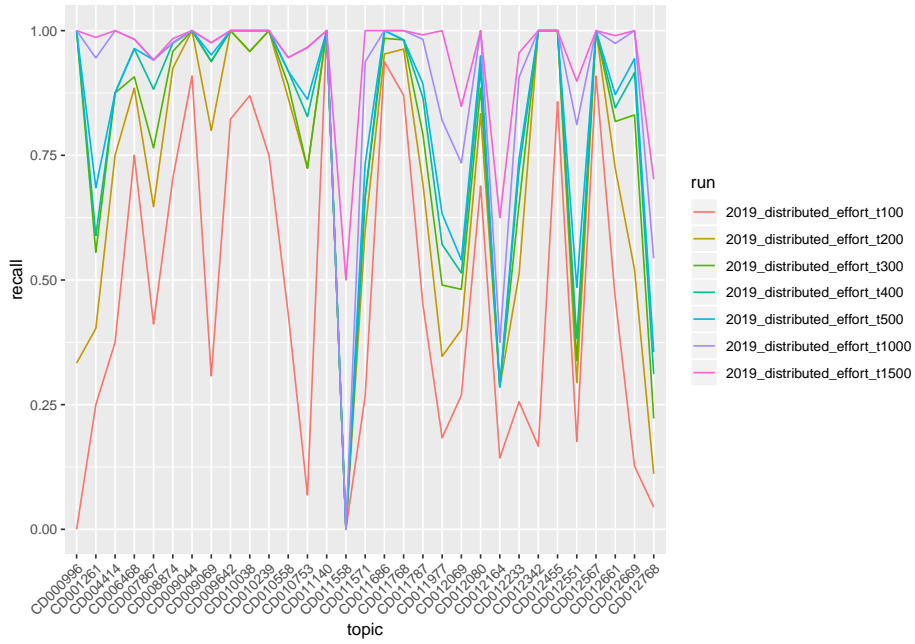
References

1. Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker, editors. *CLEF 2019 Technology Assisted Reviews in Empirical Medicine Overview. CLEF 2019 Evaluation Labs and Workshop: Online Working Notes.*, CEUR Workshop Proceedings. CEUR-WS.org, 2019.
2. Liadh Kelly, Hanna Suominen, Lorraine Goeriot, Mariana Neves, Evangelos Kanoulas, Dan Li, Leif Azzopardi, Rene Spijker, Guido Zuccon, Jimmy, and Joao Palotti, editors. *Overview of the CLEF eHealth Evaluation Lab 2019. CLEF 2019 - 10th Conference and Labs of the Evaluation Forum.* Lecture Notes in Computer Science (LNCS), Springer, September 2019.
3. Giorgio Maria Di Nunzio. A new decision to take for cost-sensitive naïve bayes classifiers. *Inf. Process. Manage.*, 50(5):653–674, 2014.
4. Giorgio Maria Di Nunzio. Interactive text categorisation: The geometry of likelihood spaces. *Studies in Computational Intelligence*, 668:13–34, 2017.
5. Giorgio Maria Di Nunzio. A study of an automatic stopping strategy for technologically assisted medical reviews. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, pages 672–677, 2018.
6. Giorgio Maria Di Nunzio, Giacomo Ciuffreda, and Federica Vezzani. Interactive sampling for systematic reviews. IMS unipd at CLEF 2018 ehealth task 2. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*, 2018.

7. Giorgio Maria Di Nunzio, Maria Maistro, and Federica Vezzani. A gamified approach to naïve bayes classification: A case study for newswires and systematic medical reviews. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1139–1146, 2018.
8. Giorgio Maria Di Nunzio, Maria Maistro, and Daniel Zilio. Gamification for machine learning: The classification game. In *Proceedings of the Third International Workshop on Gamification for Information Retrieval co-located with 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016), Pisa, Italy, July 21, 2016.*, pages 45–52, 2016.
9. Giorgio Maria Di Nunzio, Maria Maistro, and Daniel Zilio. The university of padua (IMS) at TREC 2016 total recall track. In *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016*, 2016.



(a) Topic by topic BM25 results



(b) Topic by topic distributed effort results

Fig. 1: Results for BM25 and distributed effort runs



(a) Topic by topic original 2018 p10 results



(b) Topic by topic original 2018 p50

Fig. 2: Results for original 2018 p10 and p50 runs

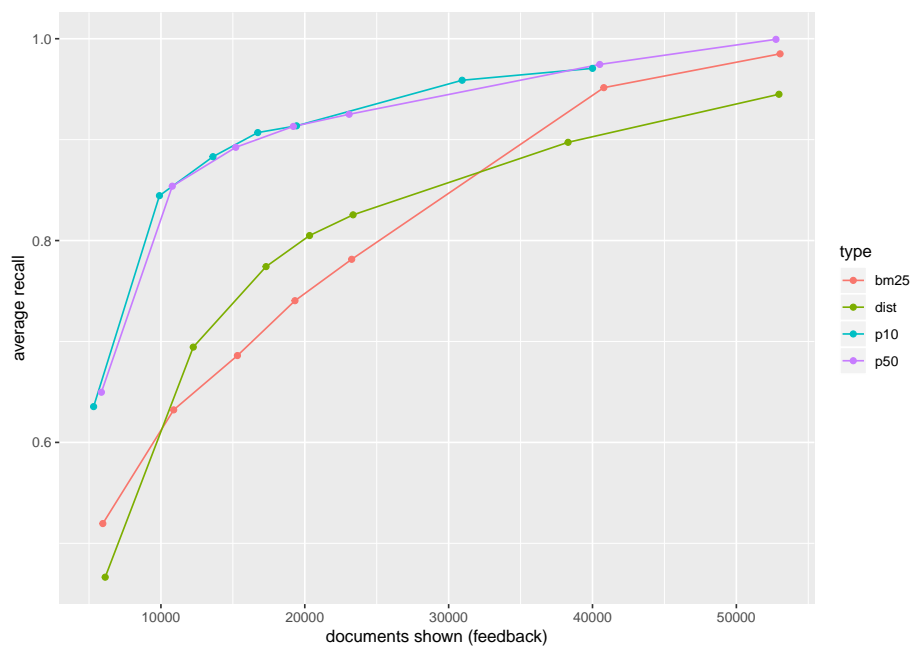


Fig. 3: Recall vs number of documents shown to the user