Combining Inter-Review Learning-to-Rank and Intra-Review Incremental Training for Title and Abstract Screening in Systematic Reviews

Antonios Anagnostou, Athanasios Lagopoulos, Grigorios Tsoumakas, and Ioannis Vlahavas

School of Informatics, Aristotle University of Thessaloniki,
54124 Thessaloniki, Greece,
{anagnoad,lathanag,greg,vlahavas}@csd.auth.gr

Abstract. We describe the approach we employed for Task II of CLEF eHealth 2017, concerning title and abstract screening in diagnostic test accuracy reviews. Our approach combines a learning-to-rank model trained across multiple reviews with a model focused on the given review, incrementally trained based on relevance feedback. Our learning-to-rank model is built using extreme gradient boosting on features computed by considering the similarity of different fields of the documents (title, abstract), with different fields of the topics (title, query). Our incrementally trained model is a support vector machine trained on a TF-IDF representation of title and abstract of the documents. The results of our approach are promising, reaching 0.658 normalized cumulative gain in the top 10 ranked documents in the simple evaluation setting and 0.846 in the cost-effective evaluation setting, the latter assuming feedback can be obtained from an intermediate user/oracle instead of the end-user.

1 Introduction

Evidence-Based Medicine (EBM) is an approach to medical practice that makes use of the current best clinical evidence in making decisions about the care and treatment of individual patients [13]. Researchers in the medical domain conduct systematic research to find the best available evidence and form review articles summarizing their discoveries on a certain topic. These *systematic reviews* usually include three stages:

1. **Document retrieval**. Experts build a Boolean query and submit it to a medical database, which returns a set of possibly relevant documents. Boolean queries typically have very complicated syntax and consist of multiple lines. Such a query can be found for reference in Listing 1.1.
2. **Title and abstract screening**. Experts go through the title and abstract of the set of documents retrieved by the previous stage and perform a first level of screening.
3. **Document screening**. Experts go through the full text of each document that passes the screening of the previous stage to decide whether it will be included in their systematic review.

Considering the rapid pace with which libraries of medical articles are expanding, Systematic Review can be a very difficult and time-consuming task.

Task II [7] [9] of CLEF eHealth 2017 lab concerns Technologically Assisted Reviews in Empirical Medicine, focusing on Diagnostic Test Accuracy (DTA), and aims to automate the second part of this process by ranking the set of documents retrieved in the first stage. Its goal is to produce an efficient ordering of the documents retrieved in the first stage, by reducing the amount of documents that experts have to go through for their reviews. This can be accomplished in two stages: by classifying documents (relevant or not) and by thresholding, ie. showing only a subset of the returned documents (the ones that are highest on the list).

Listing 1.1. Example of query in data set

```
Topic: CD009786

Title: Laparoscopy for diagnosing resectability of disease in
patients with advanced ovarian cancer

Query:
exp Ovarian Neoplasms/
Fallopian Tube Neoplasms/
((ovar* or fallopian tube*) adj5 (cancer* or tumor*
or tumour* or adenocarcinoma* or carcino* or
cystadenocarcinoma* or choriocarcinoma* or malignan*
or neoplas* or metasta* or mass or masses)).tw,ot.
(thecoma* or luteoma*).tw,ot.
1 or 2 or 3 or 4
exp Laparoscopy/
laparoscop*.tw,ot.
celioscop*.tw,ot.
peritoneoscop*.tw,ot.
abdominoscop*.tw,ot.
6 or 7 or 8 or 9 or 10
5 and 11
exp animals/ not humans.sh.
12 not 13

Pids:
        12675727
        ...
```

It is the first time this task take place and very little research is previously done on the topic. Previous approaches on this problem use an ensemble of Support Vector Machines (SVM), built over different feature spaces (documents' titles, text, etc.). [15] Other approaches use Active Learning techniques to improve results' relevance by utilizing domain experts' knowledge. [14] Finally, Learning to Rank (LTR) approaches have also been tested on biomedical data and have shown promising results. [12]

Our approaches on the task are based on binary classification methods combined with existing Learning to Rank techniques. We experimented with different classifiers and we also introduce a hybrid classification mechanism which consists of two parts: an *inter-topic* classifier, based on features computed on the training set and an *intra-topic* classifier, which is trained upon the test set documents.

2 Task overview

In CLEF eHealth 2017 Task II, participants were given a total of 20 topics with the corresponding document IDs. An example of such topic can be find in Listing 1.1. Summarizing the topics' structure, they all contain:

1. A distinct **topic id**,
2. A **topic title**,
3. An **Ovid MEDLINE** query and
4. a **set of documents' PIDs** that are returned from the query.

Similarly, documents contain the following fields:

1. A distinct **pid**,
2. A **title**,
3. The **abstract text** and
4. **Mesh headings**, based on their taxonomy

The test set comprised of topics in similar structure, summing up to a total of 30 topics.

For both the training and the test set, participants were also provided with the corresponding document relevance sheet, in which relevance was provided in the format shown in Listing 1.2, where 0 denotes negative relevance and 1 denotes positive one.

Listing 1.2. Example of query/document relevance.

```
CD010438     0  4461416     0
CD010438     0  21330915    1
CD010438     0  4576350     0
CD010438     0  20813396    0
CD010438     0  12675727    1
CD010438     0  22782135    0
...

CD011984     0  12210579    0
CD011984     0  10210123    0
CD011984     0  10210120    1
...
```

For participants' evaluation, the task defined the following metrics:

1. Area under the recall-precision curve, i.e. Average Precision (metric in task's evaluation script: **ap**)

2. Minimum number of documents returned to retrieve all R relevant documents (metric in task's evaluation script: **last_rel**) a measure for optimistic thresholding
3. Work Saved over Sampling @ Recall (metric in task's evaluation script: **wss_100, and wss_95**)

$$WSS@Recall = \frac{TN + FN}{N–(1 - Recall)}$$

4. Area under the cumulative recall curve normalized by the optimal area (metric in task's evaluation script: **norm_area**)

$$\text{optimal area} = R * N–\frac{R^2}{2}$$

5. Normalized cumulative gain @ 0% to 100% of documents shown (metric in task's evaluation script: **NCG@0 to NCG@100**)
6. Total cost uniform (metric in task's evaluation script: **total_cost_uniform**)

$$\frac{m}{R} * (N - n) * C_p$$

where:
   - N is the total number of documents in the collection
   - n is the number of documents shown to the user
   - $(N - n)$ is the number of documents not shown to the user
   - m is the number of missing relevant documents
   - C_a is the cost paid for experts/users reviewing returned documents' abstracts to determine their relevance, and
   - $C_p = 2 * C_a$
7. Total cost weighted (metric in task's evaluation script: **total_cost_weighted**)

$$\sum_{i=1}^{m} \frac{1}{2^i}(N - n) * C_p$$

8. Reliability (metric in task's evaluation script: **loss_er**) [4]

$$Reliability = loss_r + loss_e$$

where
   - $loss_r = (1 - recall)^2$ (metric in task's evaluation script: **loss_r**)
   - $loss_e = \frac{n}{R+100} * \frac{100}{N})^2$ (metric in task's evaluation script: **loss_e**)
   - $recall = \frac{n_r}{R}$ (metric in task's evaluation script: **r**) and
   - n_r is the number of relevant document found and R the total number of relevant documents

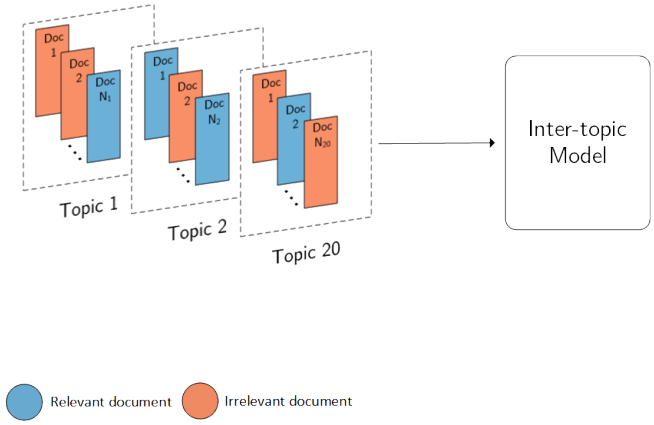

Fig. 1. Training of the inter-topic model.

3 Our Approach

The architecture of our approach, which comprises two models, is depicted in Figures 1 to 3. The first model is a learning-to-rank binary classifier that considers a topic-document pair as input and whether the document is relevant for the topic or not as output (Figure 1). This inter-topic model is used at a first stage of our approach in order to obtain an initial ranking of all documents returned by the Boolean query of an unseen test topic. The second model is a standard binary classifier that considers a document of the given test topic as input and whether this document is relevant to the test topic as output. This intra-topic model is incrementally trained based on relevance feedback that it requests after returning one or more documents to the user. The first version of this model is trained based on feedback obtained from the top k ranked documents by the inter-topic model (Figure 2). The re-ranking of subsequent documents is from then on based solely on the intra-topic model (Figure 3).

3.1 Inter-topic model

For each $\langle topic, document \rangle$ pair, we extracted a number of features, following the paradigm of [12]. The majority of the features were computed by considering the similarity of different fields of the document (title, abstract), with different fields of the topic (title, query), using a variety of similarity metrics, such as the number of common terms between the topic and the document parts, Levenshtein distance, cosine similarity or OKAPI BM25 [8]. We also computed features based solely on the topic.

In order to use the rich information available in the query field of the topics, we used Polyglot[1], a JavaScript tool that can parse and produce a full syntactical

[1] https://github.com/CREBP/sra-polyglot

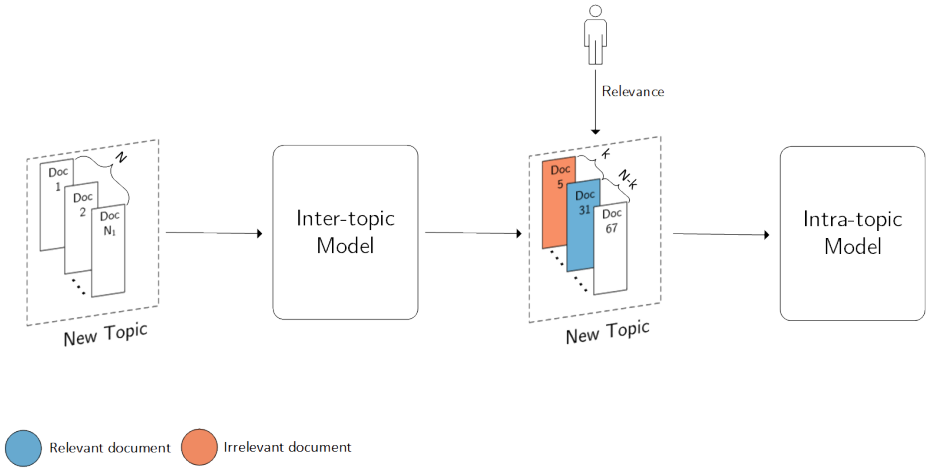

Fig. 2. Ranking with the inter-topic model. Initial training of the intra-topic model.

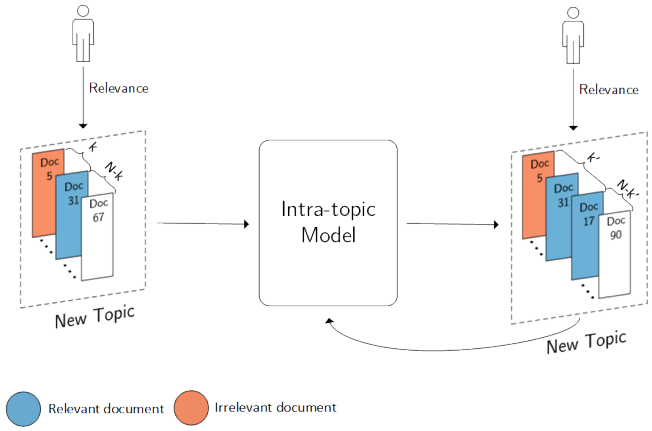

Fig. 3. Continuous re-ranking of subsequent documents and incremental re-training of the intra-topic model.

tree of Ovid MEDLINE queries. In particular, we extracted those medical subject headings (MeSH) that *should* characterize the retrieved documents, avoiding the ones that are negated in the query syntax. As an example, according to Polyglot, the MeSH terms found in the Ovid MEDLINE query of Listing 1.1 are the following:

- Ovarian neoplasms
- Fallopian Tube Neoplasms,
- Laparoscopy,
- animals (negated),
- humans

We eventually settled to the 24 features that can be found in Table 1, after extensive investigation of the performance of our model with additional variations of these features. Two of these features are only topic-dependent, denoted with T in the *Category* column of Table 1, as opposed to the rest 22 of the features that dependent on both the topic and the document, denoted with $T - D$. The notation used in the *Description* column of Table 1 is explained here:

- t represents the title of each topic, consisting of tokens t_i.
- m represent the MeSH terms extracted from the query of each topic.
- d represents the title or abstract of a document, consisting of $|d|$ tokens d_j.
- $c(x, d)$ denotes the number of occurrences of title token or MeSH term x of the topic in document d.

We have experimented with a variety of different classifiers, including Support Vector Machines [5], Gradient Boosting [6], eXtreme Gradient Boosting (XGBoost) [3] and LamdaMART [2]. The best results were achieved with XGBoost. We have also experimented with a variety of undersampling techniques, such as EasyEnsemble [10], but this did not lead to accuracy improvements.

3.2 Intra-topic model

The first version of the intra-topic model is trained based on the top k documents as ranked by the inter-topic model. We then iteratively re-rank the rest of the documents, expanding the training set of the intra-topic model with the top-ranked document, until the whole list has been added to the training set or a certain threshold is reached. This iterative feedback and reranking mechanism is described in detail in Algorithm 1. For the local classifier, a standard *TF-IDF* vectorization was used, enhanced with English stop words removal.

4 Evaluation setups and results

Task II of CLEF eHealth 2017 supported two experimental setups: one for simple evaluation and one for cost effective.

In the simple evaluation, our aim was to utilize relevance feedback as much as possible without any cap or limitation, so as to experiment with different techniques for boosting ranking metrics. In the cost-effective evaluation, we have implemented thresholding by limiting the amount of documents (column *Threshold* in Table 2) that we request feedback for and by not showing to users documents for which negative relevance was received.

In Table 3 you can find the official results for the simple evaluation setup. In Table 4 you can find the results for the cost effective evaluation, as they derive from the evaluation script provided by the task's organizers. Please note, that because of undergoing software enhancements in the script some metrics in the cost-effective evaluation might be inaccurate, e.g. the total cost metrics, as they have not be adjusted to different run outputs for that setup. Each of the runs have a parameterized version of HybridRankSVM and thresholding points, which are listed in Table 2.

ID	Description	Category	Topic field	Document field
1	$\sum_{t_i \in t \cap d} c(t_i, d)$	$T-D$	Title	Title
2	$\sum_{t_i \in t \cap d} \log c(t_i, d)$	$T-D$	Title	Title
3	$\sum_{t_i \in t \cap d} c(t_i, d)$	$T-D$	Title	Abstract
4	$\sum_{t_i \in t \cap d} \log c(t_i, d)$	$T-D$	Title	Abstract
5	$\sum_{m_i \in t \cap d} c(m_i, d)$	$T-D$	Query	Title
6	$\sum_{m_i \in t} \sum_{d_j \in d} levenshtein(m_i, d_j)$	$T-D$	Query	Title
7	$\sum_{m_i \in t} \sum_{d_j \in d} levenshtein(m_i, d_j)$ if $levenshtein(m_i, d_j) < k$	$T-D$	Query	Title
8	$\sum_{m_i \in t \cap d} \log c(m_i, d)$	$T-D$	Query	Title
9	$\sum_{m_i \in t \cap d} c(m_i, d)$	$T-D$	Query	Abstract
10	$\sum_{m_i \in t \cap d} \log c(m_i, d)$	$T-D$	Query	Abstract
11	$\sum_{m_i \in t} \log \frac{\lvert C \rvert}{df(t_i)}$	T	Query	-
12	$\sum_{m_i \in t} \log \frac{\lvert C \rvert}{df(t_i)}$	T	Query	-
13	BM25	$T-D$	Title	Title
14	BM25	$T-D$	Title	Abstract
15	BM25	$T-D$	Query	Title
16	BM25	$T-D$	Query	Abstract
17	log(BM25)	$T-D$	Title	Title
18	log(BM25)	$T-D$	Title	Abstract
29	log(BM25)	$T-D$	Query	Title
20	log(BM25)	$T-D$	Query	Abstract
21	Cosine similarity of TF-IDF representations	$T-D$	Title	Title
22	Cosine similarity of TF-IDF representations	$T-D$	Title	Abstract
23	Cosine similarity of TF-IDF representations	$T-D$	Query	Title
24	Cosine similarity of TF-IDF representations	$T-D$	Query	Abstract

Table 1. Set of features employed by our inter-topic model.

Algorithm 1: Reranking algorithm of the intra-topic model

Input : The ranked documents R, of length n, as produced by the XGBoost classifier, initial training step k, initial local training step $step_{init}$, secondary local training step $step_{secondary}$, step change threshold t_{step}, final threshold t_{final} (optional)

Output: Final ranking of documents R - $finalRanking$

$finalRanking \leftarrow ()$; `// empty list`

for $i = 1$ ***to*** k **do**

  $finalRanking_i \leftarrow R_i$

$k' \leftarrow k$;

while *not* $finalRanking$ *contains both* ***relevant*** *and* ***irrelevant*** *documents* **do**

  $k' \leftarrow k' + 1$;

  $finalRanking_{k'} = R_{k'}$;

while *not* $length(finalRanking) == n$ ***OR*** $length(finalRanking) == t_{final}$ **do**

  $train(finalRanking)$; `// Train a local classifier by asking for abstract or document relevance for these documents`

  $localRanking = rerank(R - finalRanking)$; `// Rerank the rest of the initial list` R `from the predictions of the local classifier`

  if $length(finalRanking) < t_{step}$ **then**

    $step = step_{init}$;

  else

    $step = step_{secondary}$;

  for $i = k'$ ***to*** $k' + step$ **do**

    $finalRanking_i \leftarrow localRanking_{i-k'}$;

return $finalRanking$;

Run-ID	k	$step_{initial}$	t_{step}	$step_{secondary}$	t_{final}	Threshold
1	5	1	200	100	2000	-
2	10	1	300	100	2000	-
3	10	1	200	100	1000	-
4	10	1	200	50	2000	-
5	10	1	200	100	1000	1000
6	10	1	300	100	2000	1000
7	5	1	200	100	2000	1000
8	10	1	200	50	2000	1000

Table 2. Run details for CLEF eHealth Task II

Run Id	1	2	3	4
Recall	1.0	1.0	1.0	1.0
Average Precision	0.297	0.293	0.285	0.293
wss_95	0.693	0.697	0.678	0.69
wss_100	0.519	0.521	0.511	0.519
last_rel	2143.233	2124.267	2183.267	2119.267
NCG@10	0.662	0.658	0.662	0.656
NCG@20	0.873	0.87	0.871	0.868
NCG@100	1.0	1.0	1.0	1.0
Cost(weighted)	6674.5	6677.167	5474.5	6687.833
Cost(uniform)	6674.5	6677.167	5474.5	6687.833
norm_area	0.928	0.92	0.924	0.92
loss_er	0.544	0.544	0.544	0.544

Table 3. Simple Evaluation results for CLEF eHealth Task II

Run Id	5	6	7	8
Recall	0.928	0.928	0.928	0.928
Average Precision	0.77	0.796	0.795	0.796
wss_95	0.579	0.613	0.612	0.612
wss_100	0	0	0	0
last_rel	2025	1784	1797	1775
NCG@10	0.773	0.846	0.844	0.846
NCG@20	0.901	0.932	0.931	0.932
NCG@100	0.984	0.984	0.984	0.984
Cost(weighted)	3918.7	3918.7	3918.7	3918.7
Cost(uniform)	3918.7	3918.7	3918.7	3918.7
norm_area	0.91	0.918	0.918	0.918
loss_er	0.561	0.561	0.561	0.561

Table 4. Cost-Effective Evaluation results for CLEF eHealth Task II

5 Conclusion and future work

In conclusion, in this paper we introduced a hybrid classification approach for medical document ranking. Our approach constructs a global classification model based on LTR features of the training documents, produces an initial ranking for the test documents and then iteratively asks for feedback and rerank them based on the acquired relevance.

As future work, we believe that experimentation with more features, such as semantic representations (e.g. *word2vec* [11], *LDA* [1], etc.) or different undersampling setups could boost metrics even further. Moreover, it would be worthy to experiment with other classification approaches as well, such as neural networks.

Acknowledgments

This work has been partially funded by Atypon Systems Inc.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
2. Burges, C.J.: From ranknet to lambdarank to lambdamart: An overview. Learning 11(23-581), 81 (2010)
3. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. ACM (2016)
4. Cormack, G.V., Grossman, M.R.: Engineering quality and reliability in technology-assisted review. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 75–84. SIGIR '16, ACM, New York, NY, USA (2016), http://doi.acm.org/10.1145/2911451.2911510
5. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995), http://dx.doi.org/10.1023/A:1022627411411
6. Friedman, J.H.: Stochastic gradient boosting. Computational Statistics & Data Analysis 38(4), 367–378 (2002)
7. Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: Clef 2017 ehealth evaluation lab overview. CLEF 2017 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS) (September 2017)
8. Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments: Part 2. Information processing & management 36(6), 809–840 (2000)
9. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: Clef 2017 technologically assisted reviews in empirical medicine overview. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum. CEUR Workshop Proceedings, Dublin, Ireland (2017), CEUR-WS.org

10. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory under-sampling for class-imbalance learning. In: Proceedings - IEEE International Conference on Data Mining, ICDM. pp. 965–969 (2006)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
12. Qin, T., Liu, T.Y., Xu, J., Li, H.: LETOR: A benchmark collection for research on learning to rank for information retrieval. Information Retrieval 13(4), 346–374 (2010)
13. Sackett, D.L.: Evidence-based medicine. In: Seminars in perinatology. vol. 21, pp. 3–5. Elsevier (1997)
14. Wallace, B.C., Small, K., Brodley, C.E., Trikalinos, T.A.: Active learning for biomedical citation screening. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 173–182. ACM (2010)
15. Wallace, B.C., Trikalinos, T.A., Lau, J., Brodley, C., Schmid, C.H.: Semi-automated screening of biomedical citations for systematic reviews. BMC bioinformatics 11(1), 55 (2010)