York University at CLEF eHealth 2014: A Learning-to-Rank Approach for Medical Document Retrieval

Jiajin Wu and Jimmy Huang

Information Retrieval and Knowledge Management Research Lab,
School of Information Technology,
York University
wujiajin.justin@gmail.com
jhuang@yorku.ca

Abstract. We used *learning-to-rank* methods for training ranking model. Due to the limited number of training queries, we split them and conducted 5-fold cross validation. We set the proportion of training and testing as 4 : 6. For the features used in learning the model, a total of 231 features that are from multiple information retrieval models with different parameter settings were adopted. For the baseline run, we used Random Forest method to train the models with 5-fold cross validation. Only the binary relevance information were taken into account while training the model. Five trained models from 5-fold cross validation were used on the testing data to predict scores, and then we used equal weights to linearly combine the results given by different models. For run #5, we used 8 learning to rank methods to train models separately and linearly combine them together, and the binary relevance judgment was used as well. For run #6 and #7, graded relevance were taken into consideration. The difference between run 6# and #7 is that for run #6, multiple *learning-to-rank* methods were used while for run #7, only Random Forest method was used. The best result of the four runs is achieved by run #5, which used multiple models combination based on binary relevance judgment.

1 Introduction

These working notes serve to present the experimental method presented by YorkU in the CLEF eHealth 2014 task 3a [7] which consists of retrieving relevant medical documents for the user queries. Five training queries and fifty testing queries were provided in the task. The goal of the task is to retrieve relevant documents from approximate one million medical documents for the user queries. For more details about this task and related tasks, please refer to [10]. Our main objective in performing this task is to provide a solution that requires no manual tuning of parameters. Secondly, we want to test the performance of *learning-to-rank* [11] method in medical document retrieval.

To achieve the main goal, we used supervised *learning-to-rank* method based on the provided five training queries to train the models. Due to the limitedness of training dataset, we used various strategies to combine the trained models and tested on the testing set in order to get balanced results.

2 Learning-to-Rank

Learning-to-rank is a new type of method in information retrieval (IR), which has been merged in the past decade. Different from traditional ranking models in IR, *learning-to-rank* adopts machine learning approaches to solve the ranking problem. Similarly to other machine learning methods, *learning-to-rank* methods are based on features and in most cases are supervised methods, which means labeled training data is required. One edge of this type of methods is that it saves the pain for tuning parameters which is usually time consuming and tedious in traditional IR models. In the previous study of medical IR, traditional IR models were used extensively [4], [9], but *learning-to-rank* has rarely been studied.

In this work, we used an in-house IR platform to do first pass retrieval for the training dataset. Multiple retrieval models with different parameter settings were used to retrieve relevant documents for the training queries. Based on the qrels information (relevance judgment) for the five training queries provided in the task dataset, and the retrieval results from first pass retrieval, the candidate training documents were selected. Only those documents appearing in the qrels and has more than m non-zero scores from the n retrieval results were selected. Where n stands for the total number of retrieval models accounting same retrieval model with different parameter settings as different models. n in this work is 231, and m was chosen as 180.

Table 1. Training dataset

Query	# relevant	# irrelevant
1	23	18
2	25	17
3	37	13
4	31	10
5	18	22
total	134	80

Table 1 lists the numbers of relevant/irrelevant documents provided by the dataset. As is show in this table, the available documents for training are quite limited.

3 Evaluation

3.1 Dataset

We only participated in task 3a, which is a standard TREC-style IR task using (a) the 2012 crawl of approximately one million medical documents made available by the EU-FP7 Khresmoi project[1] in plain text form which was used in CLEF eHealth 2013's Task 3 and (b) a new 2014 set of English general public queries that individuals may realistically pose based on the content of their discharge summaries. This collection contains documents covering a broad set of medical topics, and does not contain any patient information. The documents in the collection come from several online sources, including the Health On the Net organization certified websites, as well as well-known medical sites and databases (e.g. Genetics Home Reference, ClinicalTrial.gov, Diagnosia). Queries are generated from the discharge summaries used in Tasks 2.

3.2 Metric

Evaluation will focus on P@5, P@10, NDCG@5, NDCG@10, but other suitable IR evaluation measures will also be computed for the submitted runs (eg. MAP). P@N indicates the percentage of relevant documents within the top N results. NDCG [8] stands for normalized discounted cumulative gain, which is another common metric for evaluating models in information retrieval.

4 Baseline Run

As the baseline run, only title and description in the query can be used and no external resource (including discharge summary, corpora, ontology, etc) can be used. To keep it simple, we used single *learning-to-rank* model based on the binary relevance judgment to train our model. So far there are plenty of methods in the literature, and we chose to use RankLib[2], an open source *learning-to-rank* package which implements eight popular algorithms: MART [6], RankNet [2], RankBoost [5], AdaRank [14], Coordinate Ascent [12], LambdaMART [13], ListNet [3] and Random Forests [1]. So our concern is that which algorithm should be chosen as baseline method.

For this sake, we conducted five-fold cross validation using all the eight algorithms on training dataset. Table 2 lists the cross validation setting, where the number represents the id of training queries.

The model that was used as baseline method is Random Forest, which achieved the best average result over the five folds in terms of precision at 10.

5 Other Runs

As our run #5, multiple models binary relevance (MMBR) run trained models using all eight methods with the same five-fold cross validation setting as baseline

[1] http://www.khresmoi.eu/

[2] http://people.cs.umass.edu/ vdang/ranklib.html

Table 2. Five-fold cross validation

Train	Test
1,2	3,4,5
2,3	4,5,1
3,4	5,1,2
4,5	1,2,3
5,1	2,3,4

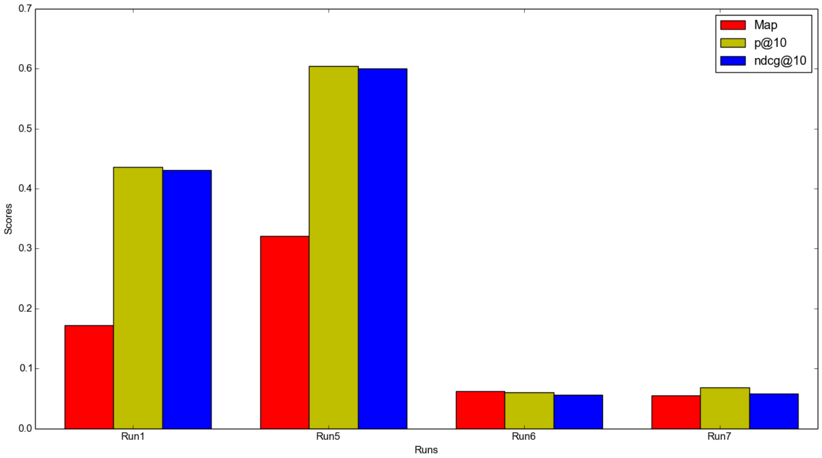

Fig. 1. Comparison of average precision, precision at 10 and normalized discounted cumulative gain at 10 results for submitted runs

run and using the binary relevance of training data. The final result on testing set is gained by combination of equal linear weights of all models. As run #6, multiple models graded relevance (MMGR) run trained models in the same way as MMBR, only different in that using graded relevance of training data. As run #7, single model graded relevance run trained model using Random Forest on five-fold cross validation. For MMBR and MMGR, the combination of multiple models are required. Since the scores range given by different models vary, the combination was conducted on top of normalization of the scores. We rescaled ranking scores to the range of $[0 - 1]$ using Formula 1,

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where X' is the normalized score, X the original score, X_{min} and X_{max} the minimal and maximal of the scores given by the model for that particular query.

The performance comparison of the four submitted runs is shown in Figure 1.

6 Discussion

As is seen in Figure 1, the best result was achieved by run #5. Notice the difference between baseline run and run #5 is that baseline run is based on single model, while run #5 uses combination of multiple models. This shows that the single model, even though it achieves the best result in training dataset, is not better than the linear equal weights combination of multiple models in the testing dataset.

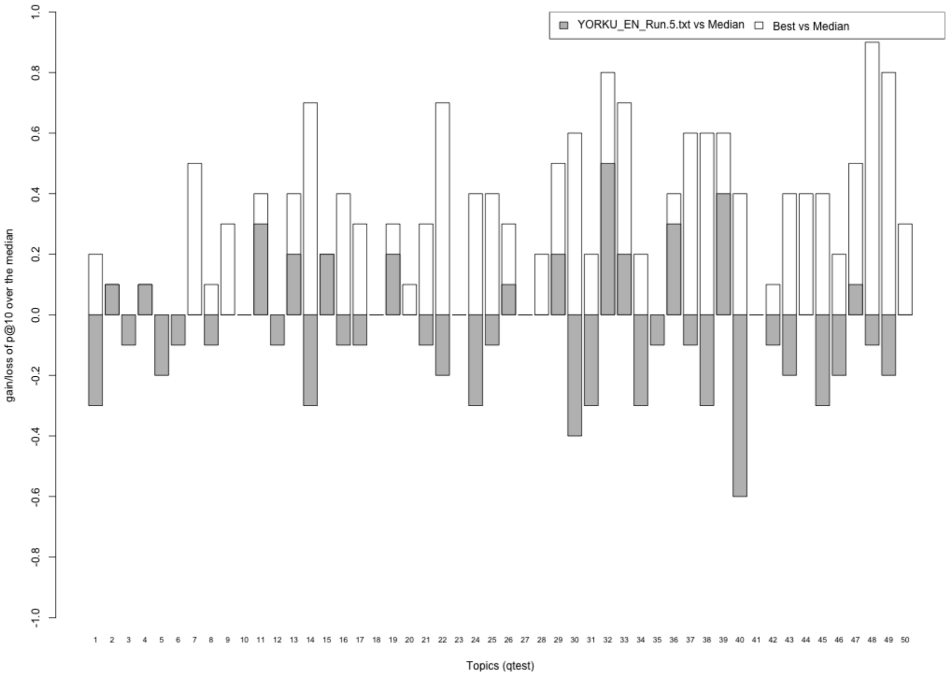

Fig. 2. Per-topic comparison between Run #5 and the other systems.

Baseline run and run #5 are relatively better than the other two runs. The main difference is that baseline run and run #5 are trained using binary relevance judgment, while run #6 and run #7 are trained using graded relevance judgment. This somewhat surprises us in that, the graded relevance provides more information to the *learning-to-rank* model about the ranking of documents, thus naturally should result in a better ranking model. But the result is contrary to this intuition. We attribute this to that the more relevance information confuses the learning models due to the shortage of training queries rather than being beneficial for the model learning.

The comparison of our best result (run #5) with the median results of all submitted runs is shown in Figure 2. It shows that, on approximate half of the queries, our best result is comparable to other systems.

7 Conclusion

In this paper, we describe our methods for medical document retrieval for task 3 in CLEF eHealth 2014. Based on supervised *learning-to-rank* methods, we have developed four strategies to conduct our experiments. The combination of multiple models using binary relevance judgment is more preferable than others. In the future, we plan to further research *learning-to-rank* in medical document retrieval, for example, 1) how domain specific features could benefit the model training, 2) how could unlabeled data be assistant in building ranking model.

Acknowledgments

This research is supported by the research grant from the Natural Sciences & Engineering Research Council (NSERC) of Canada and the Early Researcher Award/Premier's Research Excellence Award.

References

1. L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
2. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 89–96, New York, NY, USA, 2005. ACM.
3. Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 129–136, New York, NY, USA, 2007. ACM.
4. M. Daoud, D. Kasperowicz, J. Miao, and J. Huang. York university at trec 2011: Medical records track. In *TREC*, 2011.
5. Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4:933–969, Dec. 2003.
6. J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.
7. L. Goeuriot, L. Kelly, W. Li, J. Palotti, P. Pecina, G. Zuccon, A. Hanbury, G. Jones, and H. Mueller. Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval. In *Proceedings of CLEF 2014*, 2014.
8. K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002.
9. D. Kasperowicz and J. Huang. Semantic matching models for medical information retrieval: A case study. In *in the Proceedings of the 2012 Advances in Health Informatics Conference (AHIC 2012)*, 2012.

10. L. Kelly, L. Goeuriot, H. Suominen, T. Schrek, G. Leroy, D. L. Mowery, S. Velupillai, W. W. Chapman, D. Martinez, G. Zuccon, and J. Palotti. Overview of the share/clef ehealth evaluation lab 2014. In *Proceedings of CLEF 2014*, Lecture Notes in Computer Science (LNCS). Springer, 2014.
11. T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.
12. D. Metzler and W. Bruce Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007.
13. Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Inf. Retr.*, 13(3):254–270, June 2010.
14. J. Xu and H. Li. Adarank: A boosting algorithm for information retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 391–398, New York, NY, USA, 2007. ACM.