The Technion at the WWW-3 Task: Cluster-Based Document Retrieval

Fiana Raiber
Yahoo Research
fiana@verizonmedia.com

Oren Kurland
Technion
kurland@ie.technion.ac.il

ABSTRACT

Cluster-based document retrieval methods were shown to be highly effective in past research. In our submissions to the WWW-3 task, we experimented with one such method that has demonstrated superior performance compared to other state-of-the-art techniques.

TEAM NAME

Technion

SUBTASKS

English

1 INTRODUCTION

Many cluster-based document retrieval methods have been proposed over the years [5, 7, 9, 13]. Some of these methods are based on ranking clusters with respect to a query and transforming the ranking of clusters into document ranking [6, 13]. One of these methods is ClustMRF [13] that has demonstrated state-of-the-art performance: a retrieval approach based on ClustMRF was the best performing in the Web track of TREC 2013 [2, 14]. ClustMRF is a learning-to-rank approach that incorporates three types of features to represent query-cluster pairs: query-document similarities, inter-document similarities and query-independent document quality measures. Our submissions to the WWW-3 task are based on the approach taken by Raiber and Kurland [14] in TREC 2013 following its great success. Empirical evaluation reveals that our best performing run was among the top six strongest submissions [15]. We also show that the performance of this run was above the median.

2 RETRIEVAL APPROACH

We submitted five runs to the WWW-3 task as summarized in Table 1. To produce our runs, we applied a three-phase procedure. The details of the methods used in each phase are provided below.

Table 1: Summary of the five submitted runs and the collections used to train the models and to set free-parameter values.

Run	Method	Training
Technion-E-CO-NEW-1	ClustMRF [13]	ClueWeb12 Category B
Technion-E-CO-NEW-2	LTR [8]	ClueWeb12 Category B
Technion-E-CO-NEW-3	MRF [12]	Previous recommendations [12]
Technion-E-CO-NEW-4	ClustMRF [13]	ClueWeb09 Category B
Technion-E-CO-NEW-5	MEDMM [10]	ClueWeb12 Category B

Phase 1: Initial list. The Markov Random Field (MRF) method with the sequential dependence model [12] was used to retrieve an initial list of 1000 documents.

Phase 2: Learning to rank. We applied a leaning-to-rank approach [8] (LTR) to re-rank the documents in the initial list. Each query-document pair was represented using a 150-dimensional feature vector. Most of the features are based on those used in Microsoft's learning-to-rank datasets[1]. A few of the features[2] were not considered since they are not available for the document collections we use here. Following Raiber and Kurland [14], we used instead several highly effective document quality measures [1]. These features include the ratio between the number of stopwords[3] and non-stopwords in a document, the percentage of stopwords on a given stopword list that appear in the document and the entropy of the term distribution in a document. These features were computed separately for the whole document, its body, title, URL and anchor text. As an additional feature, we used the score assigned to a document by Waterloo's spam classifier [3]. To rank the documents we applied RankSVM [4] with default free-parameter values.

Phase 3: Cluster-based retrieval. The 50 highest ranked documents in the list produced in the second phase were clustered using the nearest-neighbor clustering technique [13]. The resultant 50 clusters were ranked using ClustMRF. The cluster ranking was then transformed into document ranking by replacing each cluster with its constituent documents while omitting repetitions. The order of documents within a cluster was determined based on the scores assigned to the documents by LTR in the second phase. The remaining 950 documents maintained their positions from the previous phase.

Alternative: Pseudo-relevance feedback. As an alternative to applying a cluster-based retrieval approach (phases 2 and 3), we used MEDMM [10], a state-of-the-art pseudo-feedback-based approach, to re-rank the MRF-based initial list.

3 EVALUATION

3.1 Experimental setup

Unless otherwise specified, we followed the implementation details in Raiber and Kurland [14]. To train the different models and set free-parameter values, we experimented with two approaches: (i) we used ClueWeb09 Category B with 200 topics from TREC 2009-2012 similarly to Raiber and Kurland [14], and (ii) we used ClueWeb12 Category B with 80 topics from WWW-2 [11].

[1] https://tinyurl.com/rmslr

[2] The Outlink number, SiteRank, QualityScore, QualityScore2, Query-URL click count, URL click count and URL dwell time features were not used.

[3] The stopword list includes the 100 most frequent alphanumeric terms in the collection.

Table 2: The retrieval performance of our five submitted runs. The best result in a column is boldfaced.

	NDCG@10	ERR@10	Q@10
Technion-E-CO-NEW-1	**0.658**	**0.779**	**0.682**
Technion-E-CO-NEW-2	0.656	0.750	0.674
Technion-E-CO-NEW-3	0.631	0.737	0.651
Technion-E-CO-NEW-4	0.651	0.747	0.672
Technion-E-CO-NEW-5	0.616	0.729	0.643

Table 3: The percentage of queries for which the performance of our Technion-E-CO-NEW-1 run is better than or equal to that attained by the overall best performing run (KASYS-E-CO-NEW-1) and the median performance attained by all other runs submitted to WWW-3 (Median).

	NDCG@10	ERR@10	Q@10
KASYS-E-CO-NEW-1	42.5	50.0	47.5
Median	72.5	67.1	71.3

For MRF, we set $\lambda_T = 0.85$, $\lambda_O = 0.1$ and $\lambda_U = 0.05$, following previous recommendations [12]. For MEDMM, the weight of the original query, the number of feedback terms and the number of feedback documents were selected from $\{0.1, 0.2, \ldots, 0.9\}$, $\{25, 50\}$ and $\{25, 50\}$, respectively. In addition, we set $\lambda \in \{0.1, 0.2\}$ and $\beta \in \{1.2, 1.4\}$ [10]. The free-parameter values of all the methods were selected to optimize NDCG@10. In addition to NDCG@10, we report the results for ERR@10 and Q@10 [15]. Statistically significant performance differences are determined using the two-tailed paired t-test at a 95% confidence level.

3.2 Experimental results

3.2.1 Main result. The results of our submitted runs are presented in Table 2. We can see that applying each of the three phases in our approach improved the performance: Technion-E-CO-NEW-1 (ClustMRF; third phase) outperforms Technion-E-CO-NEW-2 (LTR; second phase) which outperforms Technion-E-CO-NEW-3 (MRF; first phase). We can also see that training our models on ClueWeb12 Category B with the WWW-2 topics (Technion-E-CO-NEW-1) resulted in better performance than using ClueWeb09 Category B with the TREC 2009-2012 topics (Technion-E-CO-NEW-4). This finding suggests that training the models on the collection that is used for testing with a relatively small query set is better than training on a different collection with a larger number of queries. Indeed, of the five submissions, the highest performance in Table 2 is attained for Technion-E-CO-NEW-1 in which all three retrieval phases were applied and the ClueWeb12 Category B collection was used for training the models and optimizing free-parameter values. Finally, we can see that exploring relations between documents using a cluster-based retrieval approach resulted in better performance than utilizing a pseudo-feedback-based approach (Technion-E-CO-NEW-5). Indeed, we found that Technion-E-CO-NEW-5 is the only method in Table 2 that is statistically significantly outperformed by Technion-E-CO-NEW-1.

3.2.2 Comparison with other runs. We next compare the performance of Technion-E-CO-NEW-1, our best performing run, with that attained by the other runs submitted to WWW-3. We computed the percentage of queries for which the performance attained by Technion-E-CO-NEW-1 was higher or equal to (i) the performance attained by KASYS-E-CO-NEW-1, the best performing run among those submitted according to NDCG@10, and (ii) the median performance attained by all the other runs. Table 3 presents the results. We can see that for about half of the queries KASYS-E-CO-NEW-1 outperformed Technion-E-CO-NEW-1. Nevertheless, Technion-E-CO-NEW-1 was at least as effective as the median for about 70% of the queries suggesting that the performance of ClustMRF was above the median as in past findings [14].

4 CONCLUSIONS

In this paper we described our submissions to the WWW-3 task. We experimented with a three-phase state-of-the-art cluster-based document retrieval approach which was the best performing in the Web track of TREC 2013. We also experimented with a state-of-the-art pseudo-feedback-based method. Empirical evaluation demonstrates the effectiveness of applying a multi-phase retrieval approach.

REFERENCES

[1] Michael Bendersky, W. Bruce Croft, and Yanlei Diao. 2011. Quality-biased ranking of web documents. In *Proceedings of WSDM*. 95–104.

[2] Kevyn Collins-Thompson, Paul N. Bennett, Fernando Diaz, Charlie Clarke, and Ellen M. Voorhees. 2013. TREC 2013 Web Track Overview. In *Proceedings of TREC*.

[3] Gordon V. Cormack, Mark D. Smucker, and Charles L. A. Clarke. 2011. Efficient and effective spam filtering and re-ranking for large web datasets. *Journal of Informaltiom Retrieval* 14, 5 (2011), 441–465.

[4] Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of KDD*. 217–226.

[5] Oren Kurland. 2008. The opposite of smoothing: A language model approach to ranking query-specific document clusters. In *Proceedings of SIGIR*. 171–178.

[6] Oren Kurland and Carmel Domshlak. 2008. A rank-aggregation approach to searching for optimal query-specific clusters. In *Proceedings of SIGIR*. 547–554.

[7] Oren Kurland and Lillian Lee. 2006. Respect my authority! HITS without hyperlinks utilizing cluster-based language models. In *Proceedings of SIGIR*. 83–90.

[8] Tie-Yan Liu. 2011. *Learning to Rank for Information Retrieval.* Springer. I–XVII, 1–285 pages.

[9] Xiaoyong Liu and W. Bruce Croft. 2004. Cluster-Based Retrieval Using Language Models. In *Proceedings of SIGIR*. 186–193.

[10] Yuanhua Lv and ChengXiang Zhai. 2014. Revisiting the Divergence Minimization Feedback Model. In *Proceedings of SIGIR*. 1863–1866.

[11] Jiaxin Mao, Tetsuya Sakai, Cheng Luo, Peng Xiao, Yiqun Liu, and Zhicheng Dou. 2019. Overview of the NTCIR-14 We Want Web Task. In *Proceedings of NTCIR-14*.

[12] Donald Metzler and W. Bruce Croft. 2005. A Markov random field model for term dependencies. In *Proceedings of SIGIR*. 472–479.

[13] Fiana Raiber and Oren Kurland. 2013. Ranking document clusters using markov random fields. In *Proceedings of SIGIR*. 333–342.

[14] Fiana Raiber and Oren Kurland. 2013. The Technion at TREC 2013 Web Track: Cluster-based Document Retrieval. In *Proceedings of TREC*.

[15] Tetsuya Sakai, Sijie Tao, Zhaohao Zeng, Yukun Zheng, Jiaxin Mao, Zhumin Chu, Yiqun Liu, Maria Maistro, Zhicheng Dou, Nicola Ferro, and Ian Soboroff. 2020. Overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) Task. In *Proceedings of NTCIR-15*.