NTNUiS at the TREC 2014 Federated Web Search Track

Krisztian Balog
University of Stavanger
krisztian.balog@uis.no

Abstract: This paper describes our participation in the Federated Web Search track at TREC 2014. For the resource selection task we employ a learning-to-rank approach to combine various (instantiations of) resource ranking models. For the vertical selection task we treat the estimated collection relevance scores as binary judgements.

1 Introduction

We describe our participation in the Federated Web Search track at TREC 2014. Specifically, we took part in the *resource selection* and *vertical selection* tasks. For resource selection, our focus was on finding a way to effectively combine two principal strategies, Collection-centric (CC) and Document-centric (DC), we developed in prior work (Balog, 2014). We employ a learning-to-rank approach, where various instantiations of the CC and DC models, using different representations and relevance cutoff values, are used as features. We present our approach and results in Section 2. We base our *vertical selection* runs on the outcomes of resource selection step. Specifically, we use the estimated collection relevance scores as binary judgments, thereby essentially delegating the "selection" problem to the resource ranking component. The method and the results are described in Section 3.

2 Resource selection

In prior work, we presented two approaches to the resource selection task based on generative language modeling techniques (Balog, 2014). According to the Collection-centric (CC) model, each collection is represented as a term distribution, which is estimated from all sampled documents. The second model, Document-centric (DC), first scores individual sampled documents, then considers the top-K ranked ones to determine collection relevance. Despite its relative simplicity, the DC model delivers solid performance; at TREC 2013 it came very close to the top performing runs on all metrics (Demeester et al., 2014). We also experimented with the combination of the CC and DC strategies in our participation last year, using a linear mixture model, but it did not improve over the DC model. This year our aim is to find a way to effectively combine the CC and DC models. To this end, we employ learning-to-rank techniques.

Table 1: List of features used for resource selection.

Feature	Description
$\mathrm{DC}_{r,K}(q,c)$	$P(q\|c)$ estimated using the DC model representations: $r = \{\text{title}, \text{snippet}, \text{document}\}$ cutoff values: $K = \{10, 20, 50, 75, 100, 150, 200, 250, 300, 500, 1000\}$
$\mathrm{CC}_{r}(q,c)$	$P(q\|c)$ estimated using the CC model representations: $r = \{\text{title}, \text{snippet}\}$
$\text{snippets}(c)$	Number of snippets in the sample of c

2.1 Approach

We use the scores estimated by the CC and DC models as features. Specifically, we consider a number of different configurations, based on the type of document representation (title, snippet, page) and the cutoff value (K, only for the DC model). We refer to (Balog, 2014) for a detailed description of the CC and DC models. Additionally, we take collection size to be a feature as well (previously, it was incorporated as a prior collection probability). Table 1 lists our features (36 in total).

We employ a listwise learning-to-rank approach, LambdaMART (Wu et al., 2010). For training the machine learning model we use data from prior editions of the TREC FedWeb track.

2.2 Runs and results

We submitted the following runs:

NTNUiSrs1 Document-centric model using the entire document text (r = document) and a cutoff value of $K = 500$. This particular setting was chosen based on a (non-extensive) set of experiments performed on the FedWeb'13 collection.

Table 2: Results for our official resource selection runs. Best scores for each metric are in boldface.

Run	nDCG@20	nDCG@10	P@1	P@5
NTNUiSrs1	0.306	0.225	0.148	0.195
NTNUiSrs2	**0.348**	**0.281**	**0.206**	**0.257**
NTNUiSrs3	0.248	0.205	0.202	0.189

NTNUiSrs2 Learning-to-rank approach trained on the FedWeb'13 data set.

NTNUiSrs3 Learning-to-rank approach trained on the FedWeb'12 and '13 data sets.

Table 2 presents the results. We find that the learning-to-rank approach trained on FedWeb'13 outperforms the DC model by over 13% in terms of the official metric, nDCG@20 (NTNUiSrs2 vs. NTNUiSrs1). Interestingly, when training was done on both FedWeb'12 and '13 performance dropped substantially (NTNUiSrs3 vs. NTNUiSrs1). Discriminative learning is indeed a promising direction for this task, but further research is needed to understand how the training material should be composed.

3 Vertical selection

3.1 Approach

Our choice of method for the vertical selection task is closely tied to our resource selection approach. We assume that resource selection produces a relevance score $s(q,c)$ for each collection such that

$$s(q,c) = \begin{cases} > 0 & \text{c is relevant} \\ \leq 0 & \text{c is nonrelevant} \end{cases} \qquad (1)$$

Then, we simply select all collections that have a positive relevance score:

$$V(q) = \{c | s(q,c) > 0\}, \qquad (2)$$

where $V(q)$ denotes the set of selected verticals for query q. In a way, we delegate the "selection" problem to the resource ranking component.

3.2 Runs and results

We submitted the following runs:

NTNUiSvs2 Based on resource selection run NTNUiSrs2.

NTNUiSvs3 Based on resource selection run NTNUiSrs3.

Table 3 displays precision (P), recall (R), and F1-measure (F1) for our submitted runs. Based on these results, we make the not surprising observation that better resource selection indeed leads to better vertical selection. The scores, however, are quite low in absolute terms, which suggests that the scores produced by the resource selection approach may not satisfy the criteria that we have specified regarding the signs of collection scores (cf. Eq. 1). Perhaps the underlying resource selection step needs to be casted as a classification task as opposed to a ranking problem.

Table 3: Results for our official resource selection runs. Best scores for each metric are in boldface.

Run	P	R	F1
NTNUiSvs2	**0.157**	**0.406**	**0.205**
NTNUiSvs3	0.145	0.281	0.177

4 Conclusions

We described our participation in the TREC 2014 Federated Web Search track. For resource selection we have experimented with a discriminative learning approach for combining numerous instantiations of resource selection models. We have shown that it can outperform a competitive baseline model, but is sensitive to the choice of the underlying training material. We have used the estimated collection relevance scores, as binary judgments, to make a selection of verticals. We have found that improvements in resource selection indeed lead to improvements in vertical selection. At the same time, making a binary judgement about the relevance of a collection proves to be a difficult problem that is not necessarily best addressed as a ranking task.

5 References

Balog, K. (2014). Collection and document language models for resource selection. In *Proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013)*.

Demeester, T., Trieschnigg, D., Nguyen, D., and Hiemstra, D. (2014). Overview of the TREC 2013 federated web search track. In *Proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013)*.

Wu, Q., Burges, C. J., Svore, K. M., and Gao, J. (2010). Adapting boosting for information retrieval measures. *Inf. Retr.*, 13(3).