The University of Stavanger at the TREC 2013 Federated Web Search Track

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

Abstract: We describe the participation of the University of Stavanger in the Federated Web Search track at TREC 2013. We focus on the resource selection problem and formulate two different approaches in a probabilistic framework based on generative language modeling techniques.

1 Introduction

We address the *resource selection* task of the TREC 2013 Federated Web Search track: ranking a given set of search engines in response to an input query. Sampled search results are made available for each search engine (referred to as *collections* from now on). Building on prior research in federated search, we formulate two collection ranking strategies using a probabilistic retrieval framework based on language modeling techniques. According to one model (Collection-centric), each collection is represented as a term distribution, which is estimated from all sampled documents. Our second model (Document-centric) first ranks individual sampled documents, then aggregates their scores to determine collection relevance. We experimented with two type of representations for the sampled documents: snippets-only and full-text. Finally, we considered a linear combination of the Collection-centric and Document-centric methods.

2 Methods

We formulate the resource selection task in a generative probabilistic framework and rank collections based on their likelihood of containing documents relevant to an input query, $P(c|q)$. Instead of estimating this probability directly, we apply Bayes' rule and rewrite it to $P(c|q) \propto P(q|c)P(c)$. Thus, the score of a collection is made up of two components: (1) *query generator* ($P(q|c)$), that is, the probability of a query being generated by collection c; this can be interpreted as the collection's relevance to the query; (2) *collection prior* ($P(c)$), that is, the *a priori* probability of selecting collection c; this tells us how likely the collection is to contain the answer to any arbitrary query. We draw upon our prior work for estimating these components (Neumayer et al., 2012). Specifically, we consider two query generator models, representing two main families of collection selection strategies: lexicon-based collection selection and document-surrogate methods (Shokouhi and Si, 2011). These two approaches also bear strong resemblance to the expert finding models (Model 1 vs. Model 2) of Balog et al. (2006) and to the blog feed search models (Large vs. Small Document Models) of Elsas et al. (2008). Our collection prior is a simple one, based on collection size.

2.1 Collection-centric Model

One of the simplest approaches to resource selection is to treat each collection as a single, large document (Callan et al., 1995; Si et al., 2002). Once such a pseudo-document is generated for each collection, we can rank collections much like documents. In a language modeling setting this ranking is based on the probability of the collection generating the query. Formally:

$$P(q|c) = \prod_{t \in q} \Big\{ (1-\lambda) \big(\sum_{d \in c} P(t|d) P(d|c) \big) + \lambda P(t) \Big\}^{n(t,q)}, \quad (1)$$

where $n(t,q)$ is the number of times term t is present in the query q, $P(t|d)$ and $P(t)$ are maximum-likelihood estimates of the probability of observing term t given the document and background language models, respectively, and λ is a smoothing parameter. The background language model is estimated form all sampled documents. We assume that all documents are equally important within a given collection, therefore, we set $P(d|c)$ uniformly to $1/|c|$, where $|c|$ is the number of (sampled) documents in collection c.

2.2 Document-centric Model

Instead of creating a direct term-based representation of collections, our second approach models and queries individual sampled documents, then aggregates their relevance estimates:

$$P(q|c) = \sum_{d \in c} P(d|c) \prod_{t \in q} \big((1-\lambda) P(t|d) + \lambda P(t) \big)^{n(t,q)}, \quad (2)$$

where, as before, $P(t|d)$ and $P(t)$ and the document and background term probabilities, λ is the smoothing param-

eter, and $P(d|c)$ is the importance of the document given the collection. Additionally, we apply a relevance cut-off and consider only the top N most relevant documents in the sample index for the computation of Eq. 2. This model resembles the ReDDE collection selection algorithm (Si and Callan, 2003), but we incorporate collection size as a prior and not as part of the document score aggregation.

2.3 Combination of Methods

We also employ a linear combination of the two strategies:

$$P(q|c) = \beta P_{CC}(q|c) + (1-\beta)P_{DC}(q|c), \quad (3)$$

where P_{CC} and P_{DC} are estimated using Equations 1 and 2, respectively. For the sake of simplicity, we set β to 0.5 in our experiments.

2.4 Collection Priors

We use the number of sampled results as an approximation of collection size. Thus, we set collection priors as follows:

$$P(c) = \frac{|c|}{\sum_{c'} |c'|}. \quad (4)$$

3 Runs and Results

We considered two representations: snippet-only (S) and full-page (P). In both cases we indexed all the "visible" content. We applied only standard preprocessing steps.

We submitted three runs, all of which were automatic. All runs employ collection priors. We use the value 0.1 for the smoothing parameter λ.

`UiSP` Document-centric model based on the full page content. The relevance cut-off parameter N is set to 200.

`UiSPP` Linear combination of the Document-centric and Collection-centric models. The Document-centric model corresponds to the `UiSP` run; the Collection-centric model uses a snippet-only representation.

`UiSS` Linear combination of the Document-centric and Collection-centric models, where both use a snippet-only representation.

Table 1 displays the results. We find that the Document-centric model performs best among our submitted runs. Combining it with the Collection-centric model does not bring in any further improvements (`UiSP` vs. `UiSPP`). It is also clear that the full-page representation performs significantly better than the snippet-only one, at least for the Document-centric model (`UiSPP` vs. `UiSS`).

Table 1: Results for the resource selection task.

Run	CC	DC	NDCG@20	ERR@20
`UiSP`	-	P	0.2759	0.0200
`UiSPP`	S	P	0.2736	0.0200
`UiSS`	S	S	0.1650	0.0061

4 Conclusions

We described our participation in the TREC 2013 Federated Web Search track. Building on our earlier work (Neumayer et al., 2012) we employed two different approaches based on language modeling techniques to the resource selection task. Initial results suggest that our Document-centric model provides a competitive baseline.

5 References

Balog, K., Azzopardi, L., and de Rijke, M. (2006). Formal models for expert finding in enterprise corpora. In *Proceedings of SIGIR'06*, pages 43–50.

Callan, J. P., Lu, Z., and Croft, W. B. (1995). Searching distributed collections with inference networks. In *Proceedings of SIGIR'95*, pages 21–28.

Elsas, J. L., Arguello, J., Callan, J., and Carbonell, J. G. (2008). Retrieval and feedback models for blog feed search. In *Proceedings of SIGIR'08*, pages 347–354.

Neumayer, R., Balog, K., and Nørvåg, K. (2012). Ranking distributed knowledge repositories. In *Proceedings of TPDL'12*, pages 486–491.

Shokouhi, M. and Si, L. (2011). Federated search. *Foundations and Trends in Information Retrieval*, 5:1–102.

Si, L. and Callan, J. (2003). Relevant document distribution estimation method for resource selection. In *Proceedings of SIGIR'03*, pages 298–305.

Si, L., Jin, R., Callan, J., and Ogilvie, P. (2002). A language modeling framework for resource selection and results merging. In *Proceedings of CIKM'02*, pages 391–397.