Opinions in Federated Search: University of Lugano at TREC 2014 Federated Web Search Track

Anastasia Giachanou[1], Ilya Markov[2] and Fabio Crestani[1]

[1]*Faculty of Informatics, University of Lugano, Switzerland*
[2]*Informatics Institute, University of Amsterdam, The Netherlands*
Emails: *anastasia.giachanou@usi.ch, i.markov@uva.nl, fabio.crestani@usi.ch*

Abstract

This technical report presents the work carried out at the University of Lugano on TREC 2014 Federated Web Search track. The main motivation behind our approach is to provide better coverage of opinions that are present in federated resources. On the resource selection and vertical selection steps, we apply opinion mining to select opinionated resources/verticals given a user's query. We do this by combining relevance-based selection with lexicon-based opinion mining. On the results merging step, we diversify the final document ranking based on sentiment using the retrieval-interpolated diversification method.

Keywords: federated search, resource selection, vertical selection, results merging, sentiment diversification

1 Introduction

This paper describes the participation of the University of Lugano in collaboration with the University of Amsterdam in the TREC 2014 Federated Web Search track (FedWeb14).[1] We participated in three tasks: resource selection, vertical selection and results merging. Our aims are, first, to examine the effectiveness of opinion mining approaches for the vertical and resource selection tasks and, second, to apply sentiment diversification to the results merging task and examine if this approach can lead to better retrieval performance.

Federated search, also known as Distributed Information Retrieval (DIR), offers the means of simultaneously searching multiple information resources[2] using a single search interface and includes three phases: resource representation, resource selection and results merging [4, 10].

[1] `https://sites.google.com/site/trecfedweb`

[2] In this report, the terms *resource* and *search engine* are used interchangeably to denote a set of documents that belong to the same information source.

The goal of the FedWeb track is "to evaluate approaches to federated search at very large scale in a realistic setting, by combining the search results of existing web search engines". The FedWeb14 collection is different from a typical document collection because it consists of search results retrieved from 149 different search engines, each of which is mapped to one vertical (e.g., news, sports, kids, etc).

The FedWeb14 track focuses on three tasks: vertical selection, resource selection and results merging. Vertical selection aims to identify the subset of categories that will give the most relevant results given a user's query. The aim of the resource selection task is to identify a set of the most relevant resources given the query, while in the results merging task the retrieval results from the selected resources should be merged into a single result list.

The experiments, described in this technical report, aim to explore an important issue: the effect of considering opinions on different steps of federated search. For the resource selection task, we follow approaches that combine relevance and opinion [6]. To calculate the topical relevance of resources, we apply the widely used ReDDE resource selection method [11]. To calculate the opinionatedness of resources, we use the lexicon-based approach that counts the number of SentiWordNet terms appearing in documents of each resource [2]. For the last step, i.e., combining relevance and opinion, we use CombSUM [9].

For the results merging task, we apply sentiment diversification to produce the final result which covers different sentiments, namely positive, negative and neutral. To this end, we first retrieve documents from the top-20 resources, selected at the resource selection phase. Second, we calculate document relevance scores based on their ranks and relevance scores of corresponding resources as in [8]. Third, we apply the retrieval-interpolated framework [1] to diversify results by their sentiments.

In this year's track, organisers introduced a new task: vertical selection. In this task, participants are asked to predict relevant verticals (such as news, sports, etc.) given a user's query. For this task, we simply used the ranking of resources, produced on the resource selection phase, and the mapping between these resources and corresponding verticals to produce our results.

The rest of the report is organised as follows. In Section 2 we detail our approach for resource selection, vertical selection and results merging tasks. In Section 3 we describe our experimental setup and report results. Section 4 concludes our report.

2 Opinions in Federated Search

2.1 Resource Selection

When a user submits a query to a federated search system, resource selection aims to identify the most relevant resources that will further process the query. For the resource selection task in FedWeb14, participants are given a set of queries, a set of search engines/resources and a set of sample documents for each resource. For each query, participants are asked to return a ranked list of search engines according to their relevance to the query. Our approach to the resource selection task focuses on identifying resources that are both relevant to a query and contain opinion.

In our experiments, we apply the widely used ReDDE resource selection technique [11] to produce the ranking of the resources. In particular, for every query q we first calculate retrieval scores $s(d|q)$ for documents contained in the centralized sample index (CSI). To build CSI, we use documents sampled from 149 search engines using a set of 4000 queries (sample documents are provided by the organisers). We use the DFR_BM25 retrieval function from Terrier[3] because it showed slightly better results compared to other unsupervised retrieval approaches. Then the score of resource R is calculated as follows:

$$s(R|q) = \frac{\sum_{d \in R} s(d|q)}{m} \tag{1}$$

where m is the number of documents in CSI that were sampled from resource R.

In order to calculate the opinion score of resource R, we aggregate opinion scores of documents belonging to this resource. In particular, the opinion score of resource R is calculated as:

$$o(R) = \frac{\sum_{d \in R} o(d)}{|R|} \tag{2}$$

where $o(d)$ is the opinion score of document d and $|R|$ is the number of documents sampled from resource R. The opinion score of a document is calculated as the expected opinion score of its terms:

$$o(d) = \sum_{t \in d} o(t) p(t|d) \tag{3}$$

where $p(t|d)$ is the relative frequency of term t in document d and $o(t)$ is the sentiment of the term obtained from a pre-built lexicon. The relative frequency of term t is calculated as:

$$p(t|d) = \frac{tf(t,d)}{|d|} \tag{4}$$

where $tf(t,d)$ denotes the number of occurrences of term t in document d and $|d|$ denotes the total number of words in the document.

In order to produce the final ranking of resources, we need to combine their relevance and opinion scores. To do this, we used the CombSUM data fusion method [9]:

$$s_{final}(R|q) = s_{norm}(R|q) + o_{norm}(R) \tag{5}$$

where $s_{norm}(R|q)$ and $o_{norm}(R)$ are MinMax-normalized relevance and opinion scores of resource R respectively.

2.2 Vertical Selection

In web search, a query is associated with a set of verticals each of which focuses on specific domains (e.g., news, travel, and sports) or media types (e.g., images, videos). For the vertical selection task of FedWeb14, participants are given a set of verticals and the mapping from resources to verticals. Each search engine is associated with one category, such as web, news, travel, video, etc. In order to identify the category of a query, we use the provided mapping and our results from the resource selection task. Given those, we assume that if a search engine is selected as relevant for a given user's query, then the category (vertical) of this engine can also be a category of the query.

[3] http://terrier.org

2.3 Results Merging

Given a set of most relevant resources produced on the resource selection phase and their retrieval results, results merging aims to combine those results into a single list. The results merging task in FedWeb14 considers documents retrieved from the top-20 resources. Our approach to results merging aims to diversify the final result list to cover different sentiments, namely positive, negative and neutral. To this end, we consider both relevance and opinion scores of documents when creating the final merged list.

For each query and for each resource, the organisers provide a ranked list of documents. However, document relevance scores are not available. To approximate relevance scores $s(d|q)$ for documents from resource R we transform corresponding document ranks $r(d|q)$ as follows:

$$s(d|q) = \frac{r(d|q)}{n} s(R|q) \tag{6}$$

where n is the number of documents in the result list produced by R and $s(R|q)$ is the resource selection score of R.

For the sentiment diversification step we follow the retrieval-interpolated diversification approach [1]. More specifically, we apply an adaption of the sentiment-contribution-by-strength model (SCS). According to SCS, we first need to calculate the sentiment of each document. We do this using a lexicon-based approach and the SentiWordNet lexicon [2]. In particular, we calculate the sentiment of a document as the expected sentiment of its terms:

$$sent(d) = \sum_{t \in d} sent(t) p(t|d) \tag{7}$$

where $p(t|d)$ is the relative frequency of term t in document d (see Equation (4)) and $sent(t)$ is the dictionary sentiment of t as given by SentiWordNet. The sentiment score $sent(d)$ ranges from -1 to 1, where documents with $sent(d) \in [-1, 0)$ are considered negative in terms of opinion, with $sent(d) = 0$ – neutral and with $sent(d) \in (0, 1]$ – positive.

After calculating relevance and sentiment scores for all documents returned by selected resources, we merge these documents into a single list $\mathcal{L}$ by iteratively adding documents to the final list. Here, every next document d^* should maximize the following function:

$$d^* = \text{argmax}_d (s_{norm}(d|q) + sent'(d)) \tag{8}$$

where $s_{norm}(d|q)$ is the MinMax-normalized document relevance score and $sent'(d)$ is calculated as follows:

$$sent'(d) = |sent(d)| \prod_{\substack{d' \in \mathcal{L} \\ d' \text{of same sent.}}} (1 - |sent(d')|) \tag{9}$$

where $|\cdot|$ is the *abs* function and the product is performed over documents already added to the final list $\mathcal{L}$, which are of the same sentiment as document d. Essentially, this equation promotes documents with a high sentiment score $sent(d)$ and with sentiment, that has low probability in $\mathcal{L}$.

3 Experiments

3.1 Tasks

The TREC 2014 Federated Web Search track proposed three tasks:

- **Vertical selection:** given a query and a set of verticals, the goal of this task is to select a subset of relevant verticals. In FedWeb 2014, participants are given 24 different verticals (e.g., news, blogs, videos etc).

- **Resource selection:** given a query, a set of search engines/resources and a set of sample documents for each resource, the goal of this task is to return a ranked list of search engines according to their relevance given the query.

- **Results merging:** given a query, the top-20 resources selected on the resource selection phase and their retrieval results, the goal is to merge these results into a single list.

3.2 Experimental Setup

FedWeb14 contains a collection of search results sampled from 149 search engines obtained between April and May 2014. We used Terrier to index this collection, thus, creating CSI. For the lexicon-based opinion mining methods we tried the following opinion lexicons: AmazonKLE, SentiWordNet and MPQA. Based on the experiments with the FedWeb13 dataset[4] we decided to use SentiWordNet due to its superior performance. To calculate retrieval scores of documents in CSI, we considered the following scoring functions from Terrier: BM25, DLH13, Dirichlet Language Model and DFR_BM25. The experiments on FedWeb13 showed that DFR_BM25 produces the highest MAP, so we used this scoring functions in our runs.

We submitted three runs for the resource and vertical selection tasks. One run does not consider opinion (ULuganoDFR) while the other two runs do (ULuganoColL2 and ULuganoDocL2). In ULuganoColL2 we rerank resources considering both their relevance and opinion, while in ULuganoDocL2 documents from CSI are reranked according to their opinion before resource selection is performed.

Four runs were submitted for the results merging task. Two runs included sentiment diversification while the other two not. In the ULugDFRNoOp and ULugDFROp runs the search engine ranking was obtained from the ULuganoDFR resource selection run. The ULugFWBsNoOp and ULugFWBsOp runs exploited the baseline resource selection run provided by TREC.

The tasks were performed on a set of 50 queries provided by FedWeb14. The effectiveness of vertical selection is evaluated by standard classification metrics: precision (P), recall (R) and F-measure (F1). The resource selection task is evaluated by the normalized discounted cumulative gain (nDCG), the variant introduced in [3] and the normalized precision (nP) introduced in [5]. The main metric for the results merging is nDCG.

[4] http://snipdex.org/datasets/fedweb2013

3.3 Results

The results for the vertical selection task are reported in Table 1, for the resource selection task in Table 2 and for the results merging task in Table 3. The difference between nDCG@100 and nDCG@100_local is that the latter assumes that only the top-20 selected resources contain relevant documents.

Table 1: Results for vertical selection runs.

Run	P	R	F1
ULuganoDFR	0.117	0.983	0.197
ULuganoColL2	0.117	0.983	0.197
ULuganoDocL2	0.117	0.983	0.197

Table 2: Results for resource selection runs.

Run	nDCG@20	nP@5
ULuganoDFR	0.304	0.164
ULuganoColL2	0.297	0.158
ULuganoDocL2	0.301	0.160

Table 3: Results for results merging runs.

Run	nDCG@20	nDCG@20	nDCG@100_local
ULugDFRNoOp	0.156	0.204	0.362
ULugDFROp	0.146	0.195	0.346
ULugFWBsNoOp	0.251	0.296	0.588
ULugFWBsOp	0.224	0.273	0.545

Table 1 shows that all our approaches to vertical selection perform the same. This can be explained by the fact that we did not set any thresholds on the number of selected resources and/or verticals, so our vertical selection methods suggested a large number of verticals (on average, 17 verticals out of 24). This is the main reason for high recall and low precision of our vertical selection approaches.

Tables 2 and 3 show the results on resource selection and results merging respectively. The results show that there is no significant difference between the methods that apply opinion mining or sentiment diversification in federated search and the baselines. This was not unexpected since the topics provided by FedWeb14 are not chosen in respect of their relevance to opinionated documents. On the other side, it could be the case some topics to ask for opinionated documents even if this is not required in this track. Having this in mind, FedWeb dataset seemed appropriate for our experiments as it provides the federated environment on which we could incorporate opinions in federated search.

Previously, sentiment diversification was mainly applied to controversial topics which required opinionated documents to appear in retrieval results [7]. For such topics presenting different viewpoints is important and, therefore, sentiment diversification usually performs well [1].

To verify the above hypothesis, we applied sentiment diversification to results merging on the FedWeb13 dataset with available relevance judgements and topics' descriptions. Table 4 shows results for a subset of topics from FedWeb13. The descriptions of these topics are given in Table 5. It can be seen that these topics require documents with opinion. From the results in Table 4, we observe that our approach, which diversifies the final result list by sentiment, performs better than the baseline for these topics, proving that sentiment diversification should be used for controversial queries.

Table 4: nDCG@20 for a subset of topics from FedWeb13.

	Topics			
	7007	7084	7109	7415
Baseline(No Opinion)	0.461	0.847	0.659	0.253
Diversified By Sentiment	0.497	0.854	0.745	0.331

Table 5: Topic descriptions.

Topic	Description
7007	You are looking for a thorough text review of Howl from Allen Ginsberg.
7084	You want to read some reviews about the movie 'burn after reading'.
7109	You are in New York, and are looking for a place to eat pho.
7415	You want to know which are this year's most anticipates games.

4 Conclusions

In this paper, we described our participation in the TREC 2014 Federated Web Search track. For the resource selection and vertical selection tasks, we proposed to combine topical relevance with opinion and used a lexicon-based approach to calculate the opinionatedness of resources/verticals. For the results merging task, we used retrieval-interpolated diversification to provide a comprehensive overview of various opinions in the merged result list.

The results of our participation in FedWeb14 did not manage to support the claim that applying opinion mining and sentiment diversification to federated search can lead to a better performance. This can be explained by the fact that topics in the FedWeb14 collection were not chosen for an opinion-related task and, therefore, did not require retrieving documents with opinion. On the other hand, FedWeb13 contains few topics that ask for opinions and, therefore, our methods could improve performance for those topics. We believe, this is a promising result which requires further investigation.

Acknowledgement

This research was partially funded by grant P2T1P2_152269 and the OpiTrack project of the Swiss National Science Foundation.

References

[1] E. Aktolga and J. Allan. Sentiment diversification with different biases. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR'13)*, pages 593–602, 2013.

[2] S. Baccianella, A. Esuli, and F. Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2200–2204, 2010.

[3] 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)*, 2005.

[4] F. Crestani and I. Markov. Distributed information retrieval and applications. In *Proceedings of European Conference on Information Retrieval (ECIR'13)*, pages 865–868, 2013.

[5] T. Demeester, D. Trieschnigg, D. Nguyen, and D. Hiemstra. Overview of the TREC 2013 Federated Web Search Track. In *Proceedings of the 22nd Text REtrieval Conference (TREC 2013)*, 2013.

[6] S. Gerani, M. Carman, and F. Crestani. Aggregation methods for proximity-based opinion retrieval. *ACM Transactions on Information Systems*, 30(4):1–36, 2012.

[7] M. Kacimi and J. Gamper. Diversifying search results of controversial queries. In *Proceedings of the 20th ACM international Conference on Information and Knowledge Management (CIKM'11)*, pages 93–98, 2011.

[8] I. Markov, A. Arampatzis, and F. Crestani. On CORI results merging. In *Proceedings of European Conference on Information Retrieval (ECIR'13)*, pages 752–755, 2013.

[9] J. A. Shaw and E. A. Fox. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, pages 243–252, 1994.

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

[11] L. Si and J. Callan. Relevant document distribution estimation method for resource selection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'03)*, pages 298–305, 2003.