NovaSearch at TREC 2014 Microblog Track: Reranking with Wikipedia Page Views

Flávio Martins and João Magalhães

Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
Caparica, Portugal
flaviomartins@acm.org, jm.magalhaes@fct.unl.pt

Abstract. This paper describes our participation in the TREC 2014 Microblog real-time search task. We investigate whether page views from Wikipedia can be used successfully to estimate relevant time periods for queries. To this end, we use a recently published temporal reranking method by Efron et al. [2], which uses kernel density estimation.

1 Introduction and Task Description

In the *Temporally-Anchored Ad Hoc Retrieval task*, the user wishes to search for the most recent and relevant posts. The task can be summarized as: at time t, find tweets about topic X. Therefore, systems should favor relevant and highly informative tweets about the query topic posted before the query time. Due to the nature of microblogs, it is likely that relevance has a temporal dimension. That is, relevant tweets are likely to have been published recently, close to the time of the query. Therefore, systems should also take into account the temporality of the tweets.

Participants can access the Tweets2013 corpus by issuing text queries to a search API provided by the track. Therefore we experimented with methods to temporally rerank the list of tweets returned using the search API.

Tweets2013 corpus This collection consists of approximately 240 million tweets (statuses), collected via the Twitter streaming API by crawling the public stream sample over a two-month period: 1 February, 2013 - 31 March, 2013 (inclusive). NIST created 60 topics based on this corpus each representing a information need at a specific point in time. The assessors judged the relevance of the tweet but also considered the relevance of any URLs linked from the tweet. All assessments were conducted by NIST assessors on a three-point scale of "informativeness": not relevant, relevant and highly relevant. The primary evaluation measure is MAP, precision at rank 30 cutoff and R-prec are also reported.

2 Approach

We follow the approach proposed by Efron et al. [2]. They separated the lexical and temporal signals into two components following the views of Dakka et al. [1]. To combine these two components they propose the following log-linear model

$$log\, P_\alpha(R|D,Q) = (1-\alpha)\, log\, P(R|W_D,Q) \quad (1)$$
$$+ \alpha\, log\, P(R|T_D,Q) \quad (2)$$

where, α can be tuned. In our experiments α was fixed to 0.5.

Formally, they use a standard query-likelihood estimate for $P(R|W_D,Q)$. The probability of relevance given a timestamp and the query $P(R|T_D,Q)$ is viewed as the distribution of documents relevant to a query Q over time and thus a density f_Q exists which can be estimated. Our approach uses non-parametric kernel density estimation over the timestamps of a rank obtained using a standard retrieval method with the corpus [2] as well as over an external signal: the page views of a Wikipedia page associated with each query.

We also employ the RM3 [3] method for pseudo-relevance feedback after reranking using a temporal model, since documents from peak time periods can contain more informative terms.

3 Baselines and Official Runs

QL ranks by the scores as retrieved from the track's Tweets2013 search API. Additionally the system tries to filter two classes of tweets that are not relevant as per track guidelines. Firstly, *Twitter-style* retweets as well as *RT-style retweets* are filtered out. Secondly, tweets not in the English language, using the ldig[1] project for detection. This baseline serves as a base for the other runs below and therefore all runs filter retweets and tweets not written in English.

RM3 uses the RM3 [3] method for pseudo-relevance feedback without applying any temporal reranking. Original terms and new terms were set to the same weight. The number of feedback documents was set to 50 and feedback terms to 20. Tweet replies are filtered out from the documents set used for feedback.

NovaSearch0 estimates the density of the distribution of relevant documents using KDE over the timestamps of retrieved documents and uses it for reranking.

NovaSearch1 estimates the density of the distribution of relevant documents using KDE over the timestamps of the page views of a related Wikipedia page.

NovaSearch2 combines NovaSearch0 and NovaSearch1 runs with equal weight.

[1] https://github.com/shuyo/ldig

Table 1. TREC 2014 Microblog: Temporally-Anchored Ad Hoc Retrieval task results.

Run	MAP	R-prec	P30
Median	0.4209	0.4437	0.6315
QL	0.4268	0.4566	0.6345
RM3	0.4783	0.4872	0.6564
NovaRun0	0.4836	0.4904	0.6691
NovaRun1	0.4786	0.4851	0.6679
NovaRun2	**0.4873**	**0.4950**	**0.6709**

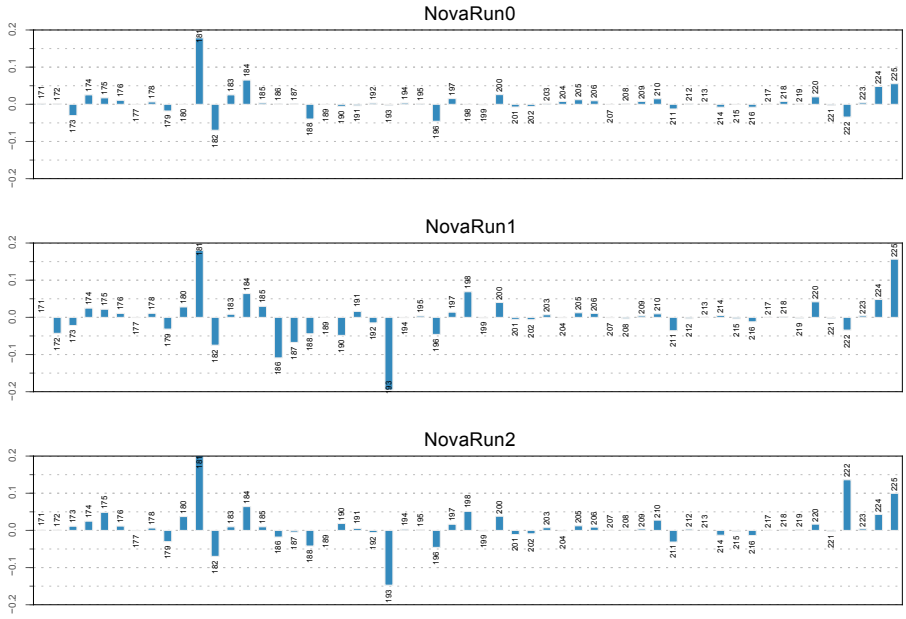

Fig. 1. Per-query differences in AP (all relevance levels) in relation to the RM3 run.

4 Summary

Our results seems to indicate that the run NovaRun2, which uses two sources of temporal evidence gives the best results. In addition, the results for NovaRun1 show that page views from Wikipedia can be used successfully to estimate relevant time periods for queries. When used in reranking, the performance improved for some queries, and deteriorated for others.

References

1. Dakka, W., Gravano, L., Ipeirotis, P.: Answering general time-sensitive queries. IEEE Transactions on Knowledge and Data Engineering 24(2), 220–235 (Feb 2012)
2. Efron, M., Lin, J., He, J., de Vries, A.: Temporal feedback for tweet search with non-parametric density estimation. In: ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 33–42. SIGIR '14 (2014), http://doi.acm.org/10.1145/2600428.2609575
3. Lavrenko, V., Croft, W.B.: Relevance based language models. In: ACM SIGIR Conference on Research and Development in Information Retrieval. p. 120127. SIGIR '01 (2001), http://doi.acm.org/10.1145/383952.383972