Query Performance Prediction and Topic Shift in Microblog Retrieval

Kuang Lu and Hui Fang

Department of Electrical and Computer Engineering
University of Delaware
140 Evans Hall, Newark, Delaware, 19716, USA
{lukuang,hfang}@udel.edu

Abstract. In Microblog retrieval, a system's ability to know when to "shut up" and how many results to return for a given query can have huge impact on its performance [1]. In addition, since relevant but redundant tweets will be deemed as irrelevant, it is also important to detect whether a tweet contain novel information or not. Therefore, in this year's Real-Time Summarization Track, we focus on estimating result cut-off thresholds and redundancy thresholds. Query performance prediction techniques [2, 3] can be used to predict the performance of a query and therefore would be helpful in deciding both thresholds. Moreover, we define topic shift of a query as the change of subtopics or aspects discussed or mentioned on Twitter over time. We suggests that using it could help us further refine the thresholds since it reflects how much new information emerges on Twitter.

1 Introduction

Real-Time Summarization Track aims to build systems to retrieve tweets for interest profiles in two scenarios: 1.tweets are posted to the evaluation broker as soon as detected relevant; 2. relevant tweets are identified at the end of each day, simulating an email digest summaries for interest profiles. In both scenarios, the retrieved tweets should be both relevant and novel.

There are two core issues that a system needs to address in order to produce good results. A study about the previous year's Microblog Track suggests that a system's ability to identify "silent days", meaning days without relevant tweets, has huge impact on the system's performance [1]. In addition, it is also important to decide how many tweets to return for each interest profile since there might not be as many as 10 relevant tweets, which is the limit per topic per day, given a query and a day. These two issues both can be solved by setting proper cut-off score thresholds. The number of returned tweets can be controlled by the score threshold. In the extreme case, if the score threshold is higher than the highest score of tweets of a day, the system will return no tweets for the day, which is ideal for a "silent day".

Query performance prediction techniques, such as clarity score [2], are used to predict the performance of a query. Intuitively, how good the performance is for a query is closely related to how many relevant documents there are at the top of the result list. Therefore, we leverage clarity scores to decide cut-off score thresholds. In addition to that, query performance predictor could help adjusting the redundancy threshold. If a query performs well in a day, the redundancy threshold can be relaxed, and vise versa. Besides query performance prediction, the topic shift of a query seems to be beneficial as well. We define topic shift of a query between two days as the change of subtopics or aspects for the query mentioned or discussed on Twitter of the two days. If the topic shift is significant for a query, it is likely that novel information about the interest profile emerges on the latter day. In this case, it is desired that our system returns more tweets on the day to cover the new subtopics.

In this year's Real-Time Summarization Track, we aim to explore how query performance prediction and topic shift could help us to accurately predict the cut-off score threshold and redundancy threshold.

2 Method Description

We use the same corpus we downloaded using twitter "sample" end point of the twitter stream api[1] for both scenarios. Non-English tweets were filtered out. For the downloaded tweets, only "id" and "text" fields are used. If a tweet is a retweet, the "text" field of the original tweet is used since retweets will be normalized to the underlying tweets for evaluation. For the interest profiles, we only use "title" field as initial queries for our system.

2.1 Scenario A

In scenario A, we mainly focus on employing query performance prediction techniques to predict the cut-off score threshold for each query of each day. Before doing that, we need a reasonable baseline to start with. Our baseline is similar to [4]. We use the "title" field of interest profiles as queries and issue these queries to three commercial search engines: Yahoo, Google, and Bing. The top 100 results' snippets of each query are retrieved and merged together to form a corpus. We use mutual information between the query terms and other terms in the corpus to select semantically related terms of the original queries and use them as expansion terms to perform retrieval on the tweet corpus [5].

After building the baseline, the next step is to predict the score thresholds using query performance prediction techniques. Clarity score [2] seems suitable for our purpose since it is a query performance predictor which indicates how ambiguous or difficult a query is. More difficult a query is, it is more likely that there are less relevant tweets of the query. However, clarity score itself is not enough since it cannot directly reflects how many relevant tweets are on the top of the result list. Considering a theoretical case when a corpus replicates itself several times, the number of relevant tweets of a query will increase if there are any relevant tweets in the original corpus, whereas the clarity score of the query will remain the same. Therefore, we incorporate the highest score of the result list, as well as the score difference between top 10 results into our system. This additional information shows the score change trend of the top scores which may indirectly show how many relevant tweet are likely to be in the top results.

We trained a linear regression model based on the idea described above to predict the score threshold. More specifically, the clarity score, the highest score, and score differences between top 10 results of one day are used as features, while the ideal score threshold of the next day is used as the expected output value of the system. When deciding the ideal score threshold for a day, we choose to use two different performance metrics: precision and f1, which results in two different regression models, which are named as precision model and f1 model. These trained models are then used to on refine the results of the baseline. Afterwards, a simple redundancy check, which uses Jaccard distances between a tweet and all previous retrieved tweets, is used to remove redundant tweets. We submitted three runs described below:

- $UDInfoSPP$: In this run, the snippets are crawled before the evaluation period started and all the expansion terms are generated based on these snippets. The precision model is used to predict score threshold.
- $UDInfoSFP$: Similar to the previous run, except that the f1 model is used.
- $UDInfoDFP$: Similar to $UDInfoDFP$, except that the snippet corpus of the new interest profiles of this year, whose topic ids have "RTS" as prefix, are crawled at the beginning of everyday. Thus, the expansion terms are newly generated each day. In this way, we hope that the expansion terms could capture the emerging new aspects of the interest profiles.

2.2 Scenario B

In scenario B, in addition to using query performance prediction, we aim to explore how the topic shifts of queries between days could help deciding the score thresholds and redundancy thresholds. If topic shift seems to be small between two days, it is reasonable to return fewer, or even no, results for the latter day, and vice versa. Similarly, when the topic shift is very significant, the redundancy threshold could be decreased in order to capture the emerging information of the query for the latter day.

[1] https://dev.twitter.com/streaming/reference/get/statuses/sample

The topic shift can be captured by measuring the difference between relevance language models between the two days. Since the true relevant language model for a query is unknown, it needs to be estimated. We chose to estimate the relevant language model of a query for a day by using top 100 results of the query for the day generated by the baseline system. However, since the accuracy of the estimation varies among days and queries, using only the estimated relevant language model differences can be misleading. If a query performs well on a day for a query, it is likely that our estimation closely captures the original model. On the other hand, if a query performs poorly on a day, or in the extreme case that there are no relevant tweets for the day, our estimation can hardly be trusted. Therefore, in order to correctly use the estimated relevant language model differences, we also need to incorporate the query performance information of the days that the relevant language models are estimated. In other words, by combining these two types of information, we could more accurately estimate the topic shift over time.

Based on the idea discussed above, we incorporate the topic shift into our cut-off score threshold framework by adding the difference between the estimated relevant language models between a day and the day before, as well as the clarity scores of the two days of the query as additional features to the linear regression model used in scenario A. We tried various language model difference measures and tested them on 2015 Microblog Track data. Jaccard distance was chosen since it produced the best performance. It is also important to note that, instead of using f1 and precision, we used ndcg@10 in scenario B for tuning since it is the official evaluation metric for scenario B.

For redundancy threshold, we trained a linear regression model based on last year's data. Given a query and any two days of the evaluation period, we select one tweet from each day to form a tweet pair. The topic shift information, meaning the clarity scores and language model difference, of these two days for the query was computed and used as features. For all the tweet pairs of the two days, the Jaccard distances are computed and the optimal value of it in terms of f1 score for redundancy classification is used as the predicted threshold.

We submitted three runs for scenario B to explore the usefulness of topic shift:

- $UDInfo\_TN$: A baseline run for scenario B without using topic shift. It is similar to *UDInfoSFP* of scenario A except the fact that we use the clarity, top score, and score differences of the day itself instead of that of the previous day to predict the cut-off score threshold of the current day. Furthermore, we use ndcg@10 as the metric in training the cut-off threshold prediction model. The redundancy threshold is the same as what is used in scenario A: a fixed redundancy threshold.

- $UDInfo\_TlmN$: It is built by adding topic shift features to the cut-off threshold prediction to $UDInfo\_TN$.

- $UDInfo\_TlmNlm$: It is built by adding topic shift features to the redundancy threshold prediction to $UDInfo\_TlmN$.

3 Results and Analysis

Table 1: Performances of Submitted Runs

(a) Performances of Task A

Run Name	$EG-1$
$UDInfoSPP$	0.0915
$UDInfoSFP$	0.0642
$UDInfoDFP$	0.0699

(b) Performances of Task B

Run Name	$nDCG@10-1$
$UDInfo\_TN$	0.1315
$UDInfo\_TlmN$	0.1451
$UDInfo\_TlmNlm$	0.1445

The performances of submitted runs for different tasks are shown in Table 1. Note that only primary metrics, which are $EG - 1$ for task A, and $nDCG@10 - 1$ are reported.

4 Conclusion

References

1. Tan, L., Roegiest, A., Lin, J., Clarke, C.L.: An exploration of evaluation metrics for mobile push notifications. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '16, New York, NY, USA, ACM (2016) 741–744
2. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '02, New York, NY, USA, ACM (2002) 299–306
3. Shtok, A., Kurland, O., Carmel, D., Raiber, F., Markovits, G.: Predicting query performance by query-drift estimation. ACM Trans. Inf. Syst. **30**(2) (May 2012) 11:1–11:35
4. Yang, P., Fang, H.: Evaluating the effectiveness of axiomatic approaches in web track. In: Proceedings of the 2013 TREC conference. (2013)
5. Fang, H., Zhai, C.: Semantic term matching in axiomatic approaches to information retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '06, New York, NY, USA, ACM (2006) 115–122