On Using Inter-document Relations in Microblog Retrieval

Jesus A. Rodriguez Perez
School of Computing Science
University of Glasgow
Glasgow, UK

Yashar Moshfeghi
School of Computing Science
University of Glasgow
Glasgow, UK

Joemon M. Jose
School of Computing Science
University of Glasgow
Glasgow, UK

{Jesus.RodriguezPerez,Yashar.Moshfeghi,Joemon.Jose}@glasgow.ac.uk

ABSTRACT

Microblog Ad-hoc retrieval has received much attention in recent years. As a result of the high vocabulary diversity of the publishing users, a mismatch is formed between the queries being formulated and the tweets representing the actual topics. In this work, we present a re-ranking approach relying on inter-document relations, which attempts to bridge this gap. Experiments with TREC's Microblog 2012 collection show that including such information in the retrieval process, statistically significantly improves retrieval effectiveness in terms of Precision and MAP, when the baseline performs well as a starting point.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval - *Information Search and Retrieval - Search Process*

General Terms: Performance, Experimentation

Keywords: Ad-hoc Retrieval, Re-ranking, Diversification, Retrieval model, Microblog

1. INTRODUCTION

Microblogging has grown in popularity in recent years, gradually transforming the way we find out about the latest events and communicate them. The most prominent microblogging service is Twitter[1], used by millions of people around the globe posting over 340 million tweets every day[2].

The most important aspect of Twitter is that it provides unique insight into events, such as first hand reports of events as they are developing, along with the opinion of those discussing them. Ad-hoc retrieval has been actively studied in the context of Twitter, in particular during the Microblog tracks at TREC 2011 and 2012 [1]. However, searching in Twitter can be extremely challenging because of document morphology and limited content. Messages posted to Twitter (known as *tweets*) are limited to 140 characters in length and consequently they are generally of a varied linguistic quality, often containing bad grammar, spelling mistakes or slang and abbreviations to overcome the length restriction. The main problem with ad-hoc retrieval of microblogs is that of the term mismatch between the query terms and those contained within the relevant tweets. To address this problem, previous works have used temporal features [2], term burstiness [3] or other approaches.

In this paper, we draw inspiration from Maximal Marginal Relevance (MMR) [4]. MMR produces a re-ranked diversified result list by taking into account how different the document being inserted is to those already on the diversified results. To this end, every document to be inserted in the re-ranked result list is compared to all documents already on it by means of cosine similarity or some other similarity measure. There are different implementations of MMR, some approaches select the document with maximum dissimilarity with respect to the documents already in the re-ranked list, whereas others compute the average dissimilarity.

Pseudo Relevance Feedback (PRF) [5] is a technique often used in conjunction with Query Expansion (QE) for the selection of relevant terms. PRF assumes that the top documents retrieved by a given query are pseudo relevant, and are therefore a good source for extracting possible relevant terms for QE.

We propose a re-ranking algorithm that promotes those tweets which are similar to a number of top documents previously selected following PRF assumptions. Furthermore, we cluster tweets treating them as a unit, for which the same re-ranking score is given. The re-ranking score is linearly combined with the initial score given by the baseline retrieval model. Our evaluation over TREC's Microblog 2012 collection shows that our re-ranking approach achieves statistically significantly better performance than the baseline, in terms of Precision and MAP evaluation metrics, when the baseline performs well as a starting point.

2. APPROACH

In this section, we introduce our approach to microblog search results re-ranking named **SimReRank**. Our approach computes a score for re-ranking, based on the similarity of document d with respect to query q and the similarity with respect to all documents in N which have been preselected by PRF. The scoring equation is given by:

$$SimReRank(d) = (1-\lambda)P(d|q) + \lambda \sum_{n \epsilon N} P(d|n) \quad (1)$$

where $P(d|q)$ is the score given by a retrieval model such as

[1] https://twitter.com/

[2] http://blog.twitter.com/2012/03/twitter-turns-six.html

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.

IDF, N is a set of documents selected following PRF principles, and $P(d|n)$ provides the similarity score of tweet d with respect to tweet n. λ represents a mixing value, linearly combining the score provided by the baseline retrieval model for document d and the similarity towards the selected top documents present in N.

In order to capture inter-document dependencies, we perform clustering of the search results given a query. The clustering is performed using a Nearest Neighbour approach as it is not limited in terms of the number of clusters that can be formed. Tweets are therefore grouped together if their similarity is higher than the threshold 0.45. This threshold was selected empirically to provide flexibility when matching similar tweets, nevertheless other thresholds should be explored in future work.

When tweets are clustered together, the centroid of the cluster is utilised instead of the tweet's term vector to be compared with the top documents in N to compute $P(d|n)$. This treats documents in a cluster as a unit, giving all tweets the same probability. Finally, a centroid is computed for each cluster averaging over the term frequencies of the tweets within.

3. EXPERIMENTAL SETUP

Dataset The evaluation has been carried out over the TREC Microblog 2012 track collection.

Metrics Following TREC evaluation procedure for comparability, we paid attention to Precision between 5 and 30 cut-off points and Mean Average Precision (MAP) at 30.

Procedure We utilise the tool Trec_eval[3] provided by TREC to compare the performance of the systems. We look at all topics together, but also at the topics for which the baseline performed best, in terms of P@5.

4. RESULTS AND DISCUSSION

Table 1 shows the percentage of relevant documents found at different cut-off points. As it is shown, the number of relevant documents decreases when we increase the cut-off point. Thus, in our experiment, we set N (defined in Section 2) to 2 since the drop in the Precision value obtained at this cut-off point compared to @1 is not significant but the probability of other tweets matching top tweets greatly increases.

Table 2 shows the performance measures for the best run of our re-ranking approach, and the baseline IDF. The best run for our re-ranking system was selected in terms of the best MAP@30, when testing the values of the mixing parameter λ in the range 0 to 1. As we can observe, the result for our re-ranking approach outperforms the baseline in average for all 60 topics in this collection. If we take into consideration all topics, the performance improvements are not statistically significant. However when ordering the topics in terms of the baseline's performance, and selecting the half of topics performing better with respect to the baseline, the performance improvements are statistically significant. This result is expected as our re-ranking approach relies on the top N tweets when promoting similar tweets up the ranking. Therefore, if those tweets are not relevant, our approach would be promoting tweets that are not relevant, resulting on a detrimental effect over those particular topics.

[3] http://trec.nist.gov/trec_eval/

Table 1: Percentage of relevant documents found at different cut-off points

Mean Relevant Docs				
@1	@2	@3	@4	@5
50.0	47.5	45.55	42.08	43.0

Table 2: Measures for the best performing run, compared against the IDF baseline. (* $p < 0.05$ and ** $p < 0.01$, when considering half of the topics with highest baseline precision.)

	Best Runs	
	IDF (baseline)	SimReRank
P@5	0.4068	**0.4237***
P@10	0.3983	**0.4153****
P@15	0.3853	**0.4045***
P@20	0.3729	**0.3847****
P@30	0.3277	**0.3542****
MAP@30	0.1091	**0.1161****

Summarising, based on these results we can confirm that our re-ranking approach provides statistically significantly better performance on average, however it does depend on the baseline retrieval model performing well when retrieving the first documents.

5. CONCLUSIONS

In this paper we have introduced an approach to re-ranking of tweets to improve ad-hoc retrieval performance. First, our approach finds relations between tweets through their clustering. Then tweets are re-scored in terms of the linear combination of their baseline scores, and their similarity with respect to the top N tweets. Experiments with TREC's Microblog 2012 collection show that our re-ranking approach statistically significantly improves retrieval effectiveness in terms of Precision and MAP evaluation metrics when the baseline performs well as a starting point.

Future work should explore the parameters utilized in depth, as well as focusing on other approaches to uncovering relevant tweet inter-relations.

6. ACKNOWLEDGMENT

This research is partially supported by the EU funded project LiMoSINe (288024).

7. REFERENCES

[1] Ounis, I., Macdonald, C., Lin, J., Soboroff, I.: Overview of the trec-2011 microblog track. In: Proceeddings of the 20th Text REtrieval Conference. (2011)

[2] Whiting, S., Moshfeghi, Y., Jose, J.M.: Exploring term temporality for pseudo-relevance feedback. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. SIGIR '11, New York, NY, USA, ACM (2011) 1245–1246

[3] Lee, C., Wu, C., Chien, T.: Burst: a dynamic term weighting scheme for mining microblogging messages. Advances in Neural Networks–ISNN 2011 (2011) 548–557

[4] Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference, ACM (1998) 335–336

[5] Cao, G., Nie, J.Y., Gao, J., Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback. In: Proceedings of the 31st annual international ACM SIGIR conference, ACM (2008) 243–250