IRIT at TREC Microblog Track 2013

Lamjed Ben Jabeur, Firas Damak,
Lynda Tamine, Guillaume Cabanac, Karen Pinel-Sauvagnat, and Mohand Boughanem

{jabeur, damak, tamine, cabanac, sauvagnat, boughanem}@irit.fr,
University Toulouse 3 - Paul Sabatier - IRIT/SIG
118 route de Narbonne F-31062 Toulouse cedex 9

Abstract. This paper describes the participation of the IRIT lab, University of Toulouse, France, to the Microblog Track of TREC 2013. Two different approaches are experimented by our team for the real-time ad-hoc search task: *(i)* a Bayesian network retrieval model for tweet search and *(ii)* a document and query expansion model for microblog search.

1 Introduction

Seeking for information over microblogs helps to find reliable, concise and real-time information about a recently happened event (few seconds ago up to few days) [1]. In contrast of Web search, where queries are submitted for informational, transactional or a navigational propose, search within microblogs differs in content, format and the underlying motivation. It is motivated by the social activity of the person as well as current events and trends that inspire microblogging community. Relevance in this context dependent on the microblogging intention and includes several factors typically social, temporal and topical relevance [2].

Microblog search is defined in TREC Microblog Track as a real-time search task is an ad-hoc retrieval task where users are interested in most recent and relevant information [3]. Results of 2011 and 2012 editions of TREC Microblog Track show that the relevance of tweets may depends on several features in addition to the textual similarity to the query such as the number of followers and followings, the freshness of information, included URLs and the user's location, etc. We investigate in this paper two different approach for microblog retrieval:

- First Bayesian network retrieval model for tweet search estimates the tweet relevance based on the microblogger influence and the temporal distribution of query terms.
- Second Document and query expansion using URLs published in tweets and Rocchio pseudo-relevance Feedback to avoid the vocabulary problem.

This paper is organized as follows. Section 2 introduces the Bayesian network model for tweet search and discusses relative results. Section 3 describes the document and query expansion model.

2 A Bayesian Network Retrieval Model for Tweet Search

Inspired from work of Pinheiro et al. [4], we propose to model tweet search using Bayesian network models that incorporate different sources of evidence into an integrated framework. As shown in figure 1, the topology of our Bayesian network model for tweet search is comprised of 3 connected networks: tweet network, microblogger network and period network. A detailed description of this model is presented in our previous paper [5].

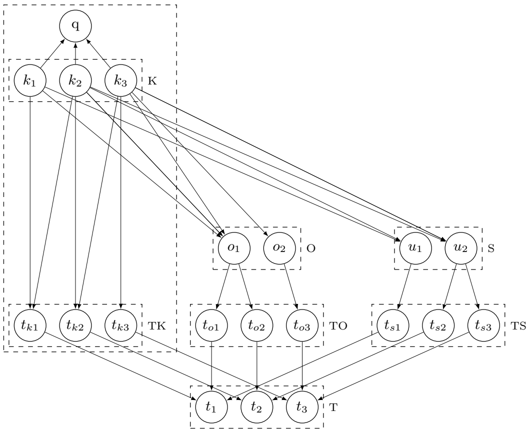

Fig. 1. Belief network model for tweet search

2.1 Query evaluation

The relevance of tweet t_j with respect to query q is computed according the topology of the Bayesian network for tweet search as follows:

$$P(t_j|q) \propto \sum_{\boldsymbol{k}} P(q|\boldsymbol{k})P(t_{kj}|\boldsymbol{k})P(t_{sj}|\boldsymbol{k})P(t_{oj}|\boldsymbol{k})P(\boldsymbol{k}) \tag{1}$$

The probability $P(t_j|\boldsymbol{k})$ that tweet t_j is generated by configuration $\boldsymbol{k}$ measures the topical similarity between the tweet and the configuration. This probability is estimated as the normalized sum of all query weights w_{k_i,t_j}.

$$w_{k_i,t_j} = \begin{cases} \frac{tf_{k_i,t_j}-\beta}{tf_{k_i,t_j}}, & \text{if } on(k_i,t_j) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where tf_{k_i,t_j} is the frequency of term k_i in tweet t_j. w_{k_i,t_j} reduces the weight of frequent terms. Accordingly, we give more importance to term presence rather than term repetition.

Assuming that the two events of observing microblogger u_f and configuration $\boldsymbol{k}$ are independent, the probability $P(t_{sj}|\boldsymbol{k})$ of observing tweet t_j having the social importance of corresponding microblogger is estimated as follows:

$$P(t_{sj}|\boldsymbol{k}) = P(t_{sj}|u_f)P(u_f) \tag{3}$$

Lets $\tau(u_f)$ be the set of tweets published by u_f. $P(t_{sj}|u_f) = \frac{1}{|\tau(u_f)|}$. Probability $P(u_f)$ is approximated to $PageRank(u_f)$, the microblogger PageRank score computed on the social network of retweets and mentions extracted from the top instantiated tweets by the query.

The probability $P(t_{oj}|\boldsymbol{k})$ of observing tweet t_j knowing period o_e and term configuration $\boldsymbol{k}$ is estimated as follows:

$$P(t_{oj}|\boldsymbol{k}) = P(t_{oj}|o_e)P(o_e|\boldsymbol{k}) \tag{4}$$

The probability $P(o_e|\boldsymbol{k})$ of selecting period o_e, having configuration $\boldsymbol{k}$, weights the different periods. We estimate this probability based on two factors. First, we consider the time decay between period o_e and query date θ_q. In fact, recent tweets are more likely to interest microblog users. Second, we consider the percentage of tweets published in o_e and containing the configuration $\boldsymbol{k}$. This highlights active period of the configuration $\boldsymbol{k}$ that concurs with a real world event. Periods are weighted as followings:

$$P(o_e|\boldsymbol{k}) = \frac{\log(\theta_q - \theta_{o_e})}{\log(\theta_q - \theta_{o_s})} \times \frac{df_{\boldsymbol{k},o_e}}{df_{\boldsymbol{k}}} \tag{5}$$

with θ_q, θ_{o_e} and θ_{o_s} are respectively the timestamps of query q, the period o_e and the period o_s when the oldest tweet containing the term configuration $\boldsymbol{k}$ is published. $df_{\boldsymbol{k},o_e}$ is the number of tweets published in o_e and containing configuration $\boldsymbol{k}$. $df_{\boldsymbol{k}}$ is the number of tweets with a term configuration $\boldsymbol{k}$.

2.2 Results and discussion

Table 1 compares results presented by different configurations of our model. *BNTSrKSO* represents our Bayesian network model with all features activated. *IRITbnetK* represents our model with only the topical feature is activated. *BNTSrKS* and *BNTSrKO* are based on the topical feature and represents our model with the social feature is activated and the temporal feature is activated, respectively. All runs were computed before relevance judgments release. First, we note that *IRITbnetK* results overpass all other configurations. Included temporal and social relevance features have not enhanced the retrieval ranking. These runs may be infected by computation problem.

	p@10	p@20	p@30	p@100	MAP
BNTSrK*	0.4617	0.4175	0.3672	0.2503	0.1952
BNTSrKO	0.3233	0.2558	0.2117	0.1283	0.0925
BNTSrKS	0.2967	0.2583	0.2222	0.1268	0.1024
BNTSrKSO*	0.2917	0.2342	0.2061	0.1278	0.1031

Table 1. Comparison of model configurations. * Official run

3 Document and Query Expansion for Microblog Search

In this approach, we propose and test several hypotheses to improve the effectiveness of microblog search engines. We mainly expanded queries and tweets since we found, by doing a failure analysis on previous track results, that the major issue of IR in microblogs was the vocabulary mismatch. This issue is mainly due to the shortness of tweets. We thus propose to extend tweets with the content of the URLs they contain, and to expand queries using Rocchio pseudo' relevance feedback [6] and Microsoft Bing spelling suggestion[1].

3.1 Design of our Approach

Our approach is composed several steps:

1. We first submitted original queries to Bing spelling suggestion API. Suggested terms were added to the original queries.
2. We then submitted the new queries to the API designed for this year's track and obtained the top 10,000 tweets for each topic.
3. We applied Rocchio pseudo-relevance feedback on resulting tweets to generate expanded queries.
4. In the next step, we submitted expanded queries to the track API, without term weights and retrieved also the top 10,000 tweets for each topic. The main goal of this step was to improve recall by enhancing our chance to get more potentially relevant tweets.
5. Then, the resulting tweets from step 2 and 4 were merged. We kept only one tweet in case of redundancy. The resulting set of tweets was composed of less than 20,000 tweets for each topic.

[1] http://www.bing.com/developers/

6. All obtained tweets were expanded with the content of the URLs they contain. We used Lucene Search engine[2] to index tweets and URLs. We specified a field for the content of tweets and a second field for the URL content published in the considered tweet, if it exists. Both fields got the same weight.
7. Considering the last set of tweets, final scores were calculated using Vector Space Model [7] and using Rocchio expansion resulting queries resulting from the step 3. We considered the term weights in queries when querying the Lucene search engine. The top-1000 high scored tweets for each topic composed our main submitted run **iritfdUrlRoc**.

We did not consider Rocchio query expansion in our second submitted run **iritfdUrl**. Only steps 1 and 2 and 6 were proceeded for it. Final scores were calculated using vector space Model and considering queries resulting from step 2.

3.2 Contribution and Experiments

Query Spelling Suggestions We noticed from previous tracks results that same named-entities are spelled differently in tweets and queries, which causes the vocabulary problem. Thus, we used Microsoft Bing spelling suggestions[3] to address this issue. This API allowed us to find the other spelled forms of a named-entity. For each query-term, we added its other spelling forms to the original query. We did not differentiate between original and added terms in terms of weighting. Although this step reformulated a noticeable number of queries (6/60) in the 2012 track, among the 60 topics of this year (2013), only 1 topic has been modified. The query "US behind Chaevez cancer" became "US behind Chaevez Chavez cancer".

Rocchio Query expansion We choose to use the improved version [6] of the original Rocchio's formula [8]. Since we considered only highly ranked documents resulting from the original queries, the Rocchio resulting formula is then:

$$Q_{new} = \alpha . Q_{orig} + \frac{\beta}{|R|} . \sum_{r \in R} r \tag{6}$$

Q_{new} is a weighted term vector for the expanded query. Q_{orig} is a weighted term vector for the original unexpanded query. R is the set of relevant documents. r is the TF/IDF term vector extracted from R. For our experiments, $\alpha = 1$ and $\beta = 0.75$ were used. The size of R was set to 10. This choice is consistent with prior experimental studies on the TREC collections. The number of expansion terms was set to 10. This choice is based on experiments done in [9] where it is found that 10 is the best number of added terms for the microblogs search task.

The aim of query expansion with Rocchio is twofold: on one hand, it could resolve the vocabulary mismatch issue, on the other hand, it addresses the problem of the importance difference among the query terms. Using Rocchio query expansion method improves results by 8% on 2012 topics compared to a run which considers original queries, with original tweets and using Vector Space Model.

Document Expansion The main observation from previous analysis on the 2012 track results, was the vocabulary mismatch, where some relevant tweets do not contain any term among the query-terms. By focusing on URLs published in the relevant tweets, we noticed that their consideration in addition to the content of tweets improve results: we obtained

[2] http://lucene.apache.org/

[3] http://www.bing.com/developers/

13.57% of improvement on 2012 test set when considering URLs using VSM and without modifying queries. The importance of URLs was also noticed in previous microblog Tracks [3, 10].

3.3 Results and Discussion

Run	Recall	Map	P@30
iritfdUrl	0.2598	0.0648	0.1394
iritfdUrlRoc*	0.2576	0.0757	0.1461

Table 2. Comparison of model configurations. * Official run

Table 2 shows results of our runs. Poor results were obtained. Additional experiments should be conducted to find the reason. The same configuration we used this year obtained $p@30 = 0.2390$ on the 2012 topics, which would have ranked us in third place among last year's automatic runs.

Query expansion did not improve recall. However, it improved the p@30 of 4.8%. The difference is significant according to Student's t test, bilateral and paired with $p < 0.05$.

References

1. Miles Efron. Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*, 62(6):996–1008, 2011.
2. Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. #twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 35–44, New York, NY, USA, 2011. ACM.
3. Iadh Ounis, Craig Macdonald, Jimmy Lin, and Ian Soboroff. Overview of the trec-2011 microblog track. In *Proceedings of the 20th Text REtrieval Conference (TREC 2011)*. NIST, 2011.
4. Marco Antnio Pinheiro de Cristo, Pvel Pereira Calado, Maria de Lourdes da Silveira, Ilmrio Silva, Richard Muntz, and Berthier Ribeiro-Neto. Bayesian belief networks for ir. *International Journal of Approximate Reasoning*, 34(2-3):163 – 179, 2003. Soft Computing Applications to Intelligent Information Retrieval on the Internet.
5. Lamjed Ben Jabeur, Lynda Tamine, and Mohand Boughanem. Intgration des facteurs temps et autorit sociale dans un modle baysien de recherche de tweets. In *Confrence francophone en Recherche d'Information et Applications (CORIA), Bordeaux, 21/03/12-23/03/12*, pages 301–316, 2012.
6. Gerard Salton and Chris Buckley. Readings in information retrieval. chapter Improving retrieval performance by relevance feedback, pages 355–364. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.
7. Doron Cohen, Einat Amitay, and David Carmel. Lucene and juru at trec 2007: 1-million queries track. In *Proceedings of the 7th Text REtrieval Conference (TREC 2007)*, page 1. NIST, 2007.
8. J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The Smart retrieval system - experiments in automatic document processing*, pages 313–323. 1971.
9. Younos Aboulnaga and Charles L.A. Clarke. Frequent Itemset Mining for Query Expansion in Microblog Ad-hoc Search. In *Proceedings of the 21th Text REtrieval Conference (TREC 2012)*. NIST, 2012.
10. Iadh Ounis, Craig Macdonald, Jimmy Lin, and Ian Soboroff. Overview of the trec-2012 microblog track. In *Proceedings of the 21th Text REtrieval Conference (TREC 2012)*. NIST, 2012.