TREC 2013 Microblog Track Experiments at Kobe University

Taiki Miyanishi, Sayaka Kitaguchi, Kazuhiro Seki, and Kuniaki Uehara

Graduate School of System Informatics Kobe University,
1-1 Rokkodai, Nada-ku, Kobe, Hyogo, 657-8501 Japan.
miyanishi@ai.cs.kobe-u.ac.jp

Abstract. This paper describes our approach to real-time ad hoc task processing in the TREC 2013 microblog track. The approach uses a concept-based query expansion method based on a temporal relevance model that uses the temporal variation of concepts (e.g. terms or phrases) on microblogs. Our model extends an effective existing word and concept-based relevance models by tracking the concept frequency over time in microblogging services. The experimentally obtained results demonstrate that our concept-based query expansion method improve search performance, especially when using tweet selection feedback.

1 Introduction

Many people search microblog documents to find temporally relevant information such as breaking news and real-time content [9], so that temporal properties (e.g., recency and temporal variations) are important factors for retrieving relevant and informative microblogs [1, 3]. Particularly, query expansion methods based on relevance feedback incorporating the temporal property of words into their models have been demonstrated as effective for improving microblog search performance [7, 8]. These time-based query expansion methods mainly use word frequency in pseudo-relevant documents as lexical information and temporal variations of word frequency as temporal information.

However, such word-based pseudo-relevance feedback (PRF) methods result in limited retrieval effectiveness for retrieving. The fundamental reason is that words have semantic ambiguity. Furthermore, word frequency often points out different time-ranges in which crowds of people are interested.

To overcome the shortcomings of word-based temporal IR, we propose a novel concept weighting scheme based on the temporal relevance model. Our approach uses concept (e.g. terms, or phrases) as with existing concept importance weighting methods [4, 5] because concepts generally have more discriminative power than words. The proposed method provides a unified framework for weighting concepts using both lexical and temporal information.

To clarify differences between the existing methods and the proposed one, Table 1 contrasts words and concepts suggested by a standard word-based PRF method [2], wTRM, a standard concept-based lexical PRF method, cTRM (Lexical) that is equal to Latent Concept Expansion (LCE) [5], and our proposed concept-based temporal PRF method using only temporal information, cTRM (Temporal), for a topic numbered MB030: *"Keith Olbermann new job"* used in the TREC microblog track. This

Table 1. Example of word and concept about a topic MB030 suggested by a word-based model (wTRM) and a concept one (cTRM).

wTRM	cTRM (Lexical)	cTRM (Temporal)
0.0642 tv	0.0132 #8(keith keith)	0.0162 #8(current tv)
0.0642 current	0.0121 #8(olbermann keith)	0.0162 #8(tv olbermann)
0.0561 countdown	0.0121 #8(olbermann olbermann)	0.0162 #8(tv keith)
0.0513 home	0.0098 #8(current keith)	0.0154 tv

topic is related to the news that Keith Theodore Olbermann who is an American sports and political commentator head for the Current TV and become the chief news officer. Table 1 clarifies that the word-based PRF method wTRM suggests topic-related words *current* and *tv*. However, *current* and *tv* often retrieve irrelevant documents at the top because these words appear in many documents. In contrast, concept-based methods cTRM (Lexical) and cTRM (Temporal) suggest exact topic-related concepts: *#uw8(olbermann keith)*, *#uw8(current tv)*, and *#8(tv olbermann)*[1]. However, cTRM (Lexical) also suggests meaningless concepts: *#8(keith keith)* and *#8(olbermann olbermann)*. Therefore, we assume that our temporal PRF method cTRM integrating lexical and temporal information for selecting topic-related concepts will be more effective than PRF method using only lexical information (e.g. LCE) as well as the standard word-based PRF method.

The remainder of the paper is organized as follows: in Section 2 we describes details of the proposed concept-based relevance model. Experimental settings and results are presented in Section 3. Finally, Section 4 presents conclusions.

2 Proposed method

Many researchers have reported recently that the concept-based IR method outperformed the word-based one across many tasks [4, 5]. However, these concept weighting approaches do not take account of temporal factors which are important factors for microblog searches. Microblog services often have real-time features by which many microblogs are posted by crowds of people when a notable event occurs. Many reports have described the effectiveness of incorporating such real-time features into PRF methods for microblog search [7, 8]. Therefore, we proposed a concept-based PRF method that combines lexical and temporal information of concepts.

We assume that the proposed concept-based relevant model $P(c|\mathcal{R})$ derives from both lexical and temporal information sources, and following the notion of bag-of-concepts, each concept c is sampled identically and independently from a lexical distribution of pseudo-relevant documents, $\mathcal{R}_l$ (top M retrieved documents), and a time distribution of ones, $\mathcal{R}_t$ (top N retrieved documents). Therefore, we have

$$\begin{aligned} P(c|Q) &= \sum_{D_l \in \mathcal{R}_l} \sum_{D_t \in \mathcal{R}_t} P(c, D_l, D_t|Q) \\ &= \sum_{D_l \in \mathcal{R}_l} \sum_{D_t \in \mathcal{R}_t} P(D_l|c, D_t, Q) P(c, D_t|Q), \end{aligned} \tag{1}$$

[1] #8uw(·) denotes an unordered window in which all words must appear within a window of 8 terms in any order.

where D_l denotes a document from pseudo-relevant documents $\mathcal{R}_l$ and D_t denotes each time (a day in our case) in $\mathcal{R}_t$. Then, we applied the simple assumption that the temporal information D_t is independent of the lexical information D_l, so that D_t is dropped from the conditional probability in Eq. 1. Therefore, we have

$$\begin{aligned} P(c|Q) &= \sum_{D_l \in \mathcal{R}_l} P(D_l|c, Q) \sum_{D_t \in \mathcal{R}_t} P(c, D_t|Q) \\ &\propto \frac{1}{P(c|Q)} \sum_{D_l \in \mathcal{R}_t} P(c, D_l|Q) \sum_{D_t \in \mathcal{R}_t} P(c, D_t|Q). \end{aligned} \tag{2}$$

Then, we assume that $P(c|Q)$ is a non-negative function and the concept c and that all query concepts are also sampled independently and identically once we choose distributions D_l and D_t. We have the score function that ranks a concept c in response to query Q as

$$S_{cTRM}(c, Q) \overset{rank}{=} \underbrace{\sum_{D_l \in \mathcal{R}_l} P(D_l)P(c|D_l) \prod_i^m P(\hat{q}_i|D_l)}_{\textbf{Lexical}} \cdot \underbrace{\sum_{D_t \in \mathcal{R}_t} P(D_t)P(c|D_t) \prod_i^m P(\hat{q}_i|D_t)}_{\textbf{Temporal}}, \tag{3}$$

where $P(c|D_l)$ is the probability of concept occurrence in a document D and $P(c|D_t)$ denotes the probability of concept occurrence at time t (day in our case). $P(\hat{q}_i|D_l)$ is the probability of a i-th query concept q_i under the concept distribution for document D. The maximum likelihood estimator of $P(\hat{q}|D)$ is $P_{ml}(c|D) = \frac{f(c;D)}{\sum_{c' \in V} f(c';D)}$. Therein, $f(c; D)$ denotes the number of concept counts of c in document D, $\sum_{c' \in V} f(c'; D)$ is the number of concepts in D where V is the set of all concepts in the vocabulary. In most cases, this probability is applied to smoothing to temper over-fitting using a given collection. Among numerous smoothing methods, the following Dirichlet smoothing [10] is often used.

$$P(c|D_l) = \frac{|D_l|}{|D_l| + \mu} P_{ml}(c|D_l) + \frac{\mu}{|D_l| + \mu} P(c|C), \tag{4}$$

where μ is the Dirichlet prior and $P(c|C)$ is a uni-gram language model of concept in a corpus C. We assume that concepts appear in a document only once because Twitter messages are very short, so that we approximate $P(c|C) \approx \frac{df(c)}{|C|}$ to avoid pre-computing concept count, where $df(c)$ is the document frequency of concept c and $|C|$ is the number of documents in a corpus. Smoothing the maximum likelihood estimator of the uni-gram language model improves the estimated probabilities.

In addition, $P(\hat{q}_i|D_t)$ is i-th query concept $\hat{q}_i$ under the concept distribution at time t. To improve our estimates for $P((\hat{q}_i|D_t)$, we also use Dirichlet smoothing as with the standard query likelihood model in Eq. 4 because the value of query likelihood $\prod_i^m P(\hat{q}_i|D_t)$ becomes 0 when a query concept $\hat{q}_i$ does not appear over time in $\mathcal{R}_t$. We have

$$P(c|D_t) = \frac{|D_t|}{|D_t| + \mu_t} \hat{P}_{ml}(c|D_t) + \frac{\mu_t}{|D_t| + \mu_t} P(c|C), \tag{5}$$

where $\hat{P}_{ml}(c|D_t) = \frac{f(c;D_t)}{\sum_{c' \in \mathcal{V}} f(c';D_t)}$, $f(c;D_t)$ is the frequency of concept c at time t, $|D_t|$ is the total number of concepts at time t, and μ_t is a parameter for smoothing.

Here $P(D_l)$ and $P(D_t)$ are uniform over all the distributions in D_l and D_t. The value of $P(c|D_t)\prod_j^{|m|} P(\hat{q}_j|D_t)$ in Eq. 3 increases when the candidate concept c and query concepts $\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_m$ were simultaneously appeared in a range. Moreover, using the probabilities of concept occurrence $P(c|D_t)$ derived from document time-stamps of pseudo-relevant documents $\mathcal{R}_t$, the PRF model derived from Eq. 3 represents real-time feature of a given topic in microblogging services. Finally, we rank candidate concepts in descending order of the association score $S_{cTRM}(c, Q)$ and use the top k concepts for query expansion.

3 Evaluation

3.1 Experimental Setup

Evaluation data. We evaluated our proposed method using the test collection for the TREC 2013 microblog track[2]. This collection consists of about 240 million tweets sampled between February 1 and March 31, 2013, for 110 search topics. To evaluate any IR system, relevance judgment is applied to the whole tweet set of each topic. The relevance levels are categorized into irrelevant (labeled 0), minimally relevant (labeled 1), and highly relevant (labeled 2). We separately evaluate our method with respect to *allrel* and *highrel* query sets: *allrel* has both minimally relevant and highly relevant tweets as relevant documents and *highrel* has only highly relevant tweets. Table 2 summarizes topic numbers that we used in our experiments.

Microblog search settings. We retrieved tweets posted before the specific time associated with each topic by Twitter tools[3] with the following setting. All queries and tweets are stemmed using the Porter stemmer without stop-word removal. They are case-insensitive.

To retrieve documents, we used a basic query likelihood model with Dirichlet smoothing [10] (we set smoothing parameter $\mu = 2500$) implemented by the Lucene search engine[4] as the language model for IR (LM) and all PRF methods used this LM as initial search results. For temporal smoothing parameter μ_t in Eq. 5, we set $\mu_t = 2500$ when retrieving documents. We retrieved 3000 tweets through Twitter API.

Then, we filtered out all non-English retrieved tweets using a language detector with infinity-gram, called ldig[5] except for a specific run, kobeU. Retweets[6] were regarded as irrelevant for evaluation in the TREC microblog track; however, we used retweets except in a final ranking of tweets because a set of retweets is a good source that might contain topic-related words for improving Twitter search performance. In accordance with the track's guidelines, all retweets including the string "RT" at the beginning of

[2] https://github.com/lintool/twitter-tools/wiki/TREC-2013-Track-Guidelines
[3] https://github.com/lintool/twitter-tools
[4] http://lucene.apache.org/core/
[5] https://github.com/shuyo/ldig
[6] Tweets re-posted by another user to share information with other users

Table 2. Summary of TREC collections and topics used for evaluation.

Name	Type	#Topics	Topic Numbers
TREC 2013	*allrel*	60	111-170
	highrel	56	111,112, 114-124, 126, 127, 129-148, 150-170

the tweet were removed from the final ranking. Finally, we used the top 1000 results for evaluation.

3.2 IR Models

We introduce the setting of the proposed PRF method based on a concept-based temporal relevance model. The concept-based method uses the combination of one or two words as a candidate concept. All concepts are extracted from tweets based on full dependence, which assumes that dependence exists between all query terms [4]. We denote the proposed PRF method combining lexical and temporal information of concepts as cTRM.

Moreover, to assess the effectiveness of incorporating concept into the retrieval model, we also proposed a word-based temporal relevance model, wTRM, that incorporates lexical and temporal information of words into its relevance model. wTRM uses only a single word as a concept in Eq. 3: wTRM does not consider multi-term concepts that combine more than two words.

For some runs, we used a two-stage relevance feedback approach which conducts PRF after relevance feedback with manual tweet selection called Tweet Selection Feedback; TSF. Miyanishi et al. [6, 8] report that this two-stage relevance feedback approaches considerably improve search result relevance for microblog search. We manually select relevant tweets from initial search results among top L tweets for TSF by each topic. We set L to 30. If relevant tweets do not exist among initial search results, we use the original user query for tweet selection feedback. All selected tweets were stopped using common stop words list with URL and mention (e.g. @trecmicroblog) removal. In the new query, the selected tweet and the original query were weighted as $1:1$ for each method using TSF. After tweet selection feedback, we conduct the proposed query expansion method using concept-based temporal relevance model, cTRM.

For query expansion methods based on PRF, we select candidate words or concepts among the top M tweets retrieved using the original query after removing the uniform resource locators (URLs), and user names starting with '@' or special characters (!, @, #, ', ", etc.). All query terms, candidates of words and concepts, and tweets are decapitalized. The candidates of words and concepts include no common stop-words. Then, we select k words or concepts among candidates in descending order of the concept weighting score such as $S_{cTRM}(c, Q)$. We use the normalized score for concept weighting. For example, the weight of i-th concept is $s_i = \frac{S_{cTRM}(c_i,Q)}{\sum_j^k S_{cTRM}(c_j,Q)}$ when using cTRM. Finally, we combined the expanded concepts of PRF with their weight and the original query as an expanded query (and selected tweets if any exists). They were weighted with $1:1$. Fig. 1 shows the example of query expansion we used. In this figure, "*water shortage*" is an original query, "*water shortage defunct bore wells ...*" is a single

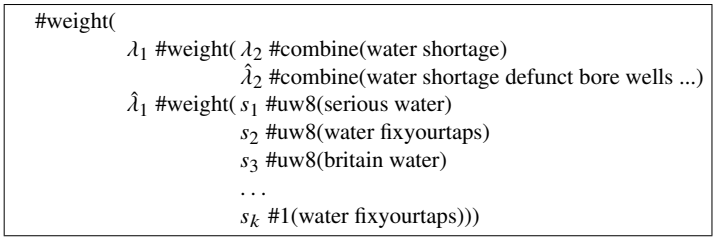

Fig. 1. Example of query expansion of topic "*water shortage*" from TREC microblog track queries.

Table 3. Descriptions of IR models

Method	Description
LM	Query likelihood model with Dirichlet smoothing
wTRM	Word-based temporal pseudo-relevance feedback
cTRM	Concept-based temporal pseudo-relevance feedback
TSF	Tweet selection feedback
URL	URL filtering

relevant tweet, and others are expand concepts. In our study, we let $\lambda_1, \hat{\lambda}_1 = 0.5$ and $\lambda_2, \hat{\lambda}_2 = 0.5$ when using TSF; $\lambda_2 = 1$ and $\hat{\lambda}_2 = 0$, otherwise.

For all query expansion methods, we tuned the parameters the number of pseudo-relevance feedback documents (i.e. M), the number of pseudo-relevant documents as temporal information (i.e. N), and the number of expansion words (i.e. k). Values of the these parameters are optimized for best performance of precision at 30 on training data. For example, we tune parameters of the IR model using TREC 2011 and 2012 microblog track dataset. Finally, we presents the description of our methods in Table 3.

3.3 Evaluation Measure

To evaluate retrieval effectiveness, we used precision at 30 (P@30) and average precision (AP). P@30 was the official microblog track metric in 2011 and both P@30 and AP were used in TREC 2012. Moreover, in TREC 2012 microblog track, "highly relevant" tweets are the required level of relevance.

3.4 Experimental Results

We summarized the results of our experiments in Table 4, which shows that kobeRMU that uses our proposed temporal PRF method outperformed kobeU for both *allrel* and *highrel* data sets in P@30 that is an objective measure for the parameter estimation. Table 4 also shows kobeTSFRM that uses TSF and the concept-based temporal query expansion method cTRM performs better than kobeTSFRMU that applies URL filtering to the results of kobeTSFRM when retrieving *allrel* documents. In contrast, kobeTSFRMU outperformed kobeTSFRM when retrieving *highrel* documents. These results suggest that URL filtering is harmful for retrieving relevant tweets; however it is effective for retrieving highly relevant and informative tweets. Moreover, kobeTSFRMU markedly outperformed kobeRMU, suggesting that relevance feedback using

Table 4. Overall results in our official runs. The best results per column are marked in boldface.

			allrel		*highrel*	
Run name	**Type**	**Method**	**AP**	**P@30**	**AP**	**P@30**
kobeU	Automatic	LM + URL - LANG	0.2217	0.4211	0.2010	0.2417
kobeRMU		wTRM + URL	0.2125	0.4311	0.1928	0.2643
kobeTSFRM	Manual	TSF + cTRM	**0.2640**	**0.4861**	0.2036	0.2720
kobeTSFRMU		TSF + cTRM + URL	0.2365	0.4733	**0.2163**	**0.2833**

manually selected a single relevant tweet as expansion words is effective to improves PRF performance further.

4 Conclusion

This paper has presented a concept-based query expansion method based on a temporal pseudo-relevance feedback (PRF) model. Our experimentally obtained results on a datasets used in TREC 20013 microblog track demonstrate that incorporating temporal information of concepts into the query expansion method improves retrieval performance. We demonstrated that using our temporal PRF method combined with URL filtering can be useful for retrieving highly relevant documents. Moreover, our two-stage relevance feedback that consists of tweet selection feedback and temporal PRF considerably improved retrieval performance for microblog search.

References

1. Efron, M., Organisciak, P., Fenlon, K.: Improving retrieval of short texts through document expansion. In: SIGIR. (2012) 911–920
2. Lavrenko, V., Croft, W.B.: Relevance based language models. In: SIGIR. (2001) 120–127
3. Lin, J., Efron, M.: Temporal relevance profiles for tweet search. In: TAIA. (2013)
4. Metzler, D., Croft, W.B.: A Markov random field model for term dependencies. In: SIGIR. (2005) 472–479
5. Metzler, D., Croft, W.B.: Latent concept expansion using Markov random fields. In: SIGIR. (2007) 311–318
6. Miyanishi, T., Seki, K., Uehara, K.: TREC 2012 microblog track experiments at Kobe university. In: TREC. (2012)
7. Miyanishi, T., Seki, K., Uehara, K.: Combining recency and topic-dependent temporal variation for microblog search. In: ECIR. (2013) 331–343
8. Miyanishi, T., Seki, K., Uehara, K.: Improving pseudo-relevance feedback via tweet selection. In: CIKM. (2013) 439–448
9. Teevan, J., Ramage, D., Morris, M.: #TwitterSearch: a comparison of microblog search and web search. In: WSDM. (2011) 35–44
10. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. TOIS **22**(2) (2004) 179–214