Feedback Model for Microblog Retrieval

Ziqi Wang and Ming Zhang(✉)

School of EECS, Peking University, Beijing 100871, China
wangziqi@pku.edu.cn, mzhang@net.pku.edu.cn

Abstract. Information searching in microblog services has become common and necessary for social networking. However, microblog retrieval is particularly challenging compared to web page retrieval because of serious vocabulary mismatch problem and non-uniform temporal distribution of relevant documents. In this paper, we propose a feedback model, which includes a feedback language model and a query expansion model considering both lexical expansions and temporal expansions. Experiments on TREC data sets have shown that our proposed model improves search effectiveness over standard baselines, lexical only expansion model and temporal only retrieval model.

Keywords: Microblog retrieval · Feedback model · Query expansion · Pseudo-relevance feedback

1 Introduction

Microblog services, such as Twitter, have become new sources of information. To get relevant information of trends or breaking news, users submit queries on the microblog sites instead of web search engines. However, microblog retrieval differs from general information retrieval (IR) due to the following reasons: (1) Tweets are short. Vocabulary mismatch problem is extremely significant in microblog retrieval. (2) Time plays an important role. Temporal distribution of relevance documents is not uniform. In this paper, we propose a novel feedback model incorporating a feedback language model and a query expansion model to tackle these challenges.

Query document vocabulary mismatch happens when user and authors of documents use different terms to represent the same concept. Vocabulary mismatch has always been a critical challenge in information retrieval. For web search, documents are relatively long and authors usually use keywords repeatedly to describe the topic. Term frequency is heavily relied on in most of retrieval models such as query likelihood model. However, in microblog retrieval, tweets have fewer terms (no more than 140 characters). Most terms, especially key concepts, only appear once in documents, which makes statistical method less reliable. Vocabulary mismatch problem becomes worse in microblog retrieval.

Since temporal distribution of relevant documents in microblog retrieval is not uniform, some work has been focusing on incorporating time information

M. Renz et al. (Eds.): DASFAA 2015, Part I, LNCS 9049, pp. 529–544, 2015.
DOI: 10.1007/978-3-319-18120-2_31

into the retrieval model. Many researchers proposed various methods of using temporal information to improve term selection in query expansion model [1] [2] [3]. Temporal evidence has also been explored under the language model framework to improve document ranking [4] [5]. It is very important to make use of temporal information.

In this paper, we propose a feedback model for microblog retrieval. Our model includes a feedback language model and a query expansion model. The feedback language model is built on the search results from the initial retrieval. Document relevance scores are adjusted based on the feedback language model. The query expansion model expands the query by using both lexical expansions and temporal expansions.

We evaluated the proposed model using the TREC 2011 and 2012 Microblog data set. The experiment results have shown that our proposed feedback language model outperforms the query likelihood baseline and our proposed query expansion model performs better than the relevance model. Overall, the proposed feedback model improves microblog retrieval effectiveness over previously proposed baselines.

The rest of the paper is organized as follows. In Section 2, we review related work including temporal information retrieval, microblog retrieval and pseudo-relevance feedback. In Section 3, we present motivation of this study. Our proposed feedback language model, query expansion model and feedback model are presented in Section 4. Experiments and analysis of results are shown in Section 5. Finally, we conclude this paper in Section 6.

2 Related Work

Related work can be found in three areas. The first is temporal information retrieval. Time plays a very important role in microblog retrieval. The second is general microblog retrieval. There has been some research focusing on other aspects besides using temporal information to improve search performance. The third is pseudo-relevance feedback via query expansion.

2.1 Temporal Information Retrieval

Previous researches incorporate recency into retrieval. Newly published documents are assumed to have a larger probability to be relevant than older documents. Li and Croft proposed a time-based language model by adding document prior based on recency [4]. Efron and Golovchinsky proposed an extension by using query-specific information to estimate parameters [1]. Massoudi et al. expanded queries by using terms in the most recent documents [6].

Instead of focusing on recency queries, some works have been trying to deal with more general time-sensitive queries. Jones and Diaz proposed a temporal query model and an approach to distinguish different types of temporal queries [7]. Dakka et al. proposed a general framework to combine lexical and temporal

evidence together [5]. Liang et al. detected burst and aggregated ranking results from different retrieval methods [8].

Pseudo-relevance feedback via query expansion has been widely used in temporal retrieval. Liang et al. proposed a two-stage pseudo-relevance feedback query expansion method [3]. Whiting et al. proposed a pseudo-relevance feedback model using the correlation between temporal profiles of n-grams obtained from query and feedback documents [9]. Whiting et al. built a graph using temporal and TF evidence and selected n-gram using PageRank [10]. Keikha et al. proposed a time-based relevance model using temporal distribution of retweets [2]. Miyanishi extended latent concept expansion model based on the temporal relevance model for query expansion [11]. Metzler et al. proposed a temporal query expansion model for event retrieval based on temporal co-occurrence of terms in a timespan [12].

The major difference between previous work and our work is how we use temporal information. Most previous work used temporal information to select lexical expansions. Our proposed model identifies bursts and conducts temporal expansions for the query. Temporal expansions and lexical expansions are then combined together in a query expansion model. Besides, our proposed model can deal with both temporally unambiguous and ambiguous queries while previous temporal models could only handle temporally unambiguous queries.

2.2 Microblog Retrieval

Query document vocabulary mismatch problem is one of the critical challenges of information retrieval, especially microblog retrieval due to the short length of documents. Methods such as query expansion and document expansion have been studied to address the query document vocabulary mismatch problem.

For query expansion methods, besides using temporal information as we discussed above, external sources can also be useful. Chen et al. used external knowledge including Google and Wikipedia to conduct query expansion [13]. Bandyopadhyay et al. proposed a query expansion model using Google API and BBC site [14].

Efforts have also been made to explore document expansion methods. Efron et al. proposed an aggressive document expansion based on pseudo-relevance feedback [15]. Han et al. proposed a document expansion by using nearest neighbors of documents [16].

One of the most important differences between microblog retrieval and web page retrieval is that many tweets are low-quality and contain a lot of noise. Choi et al. proposed a quality model to demote uninformative content [17]. Gurini and Gasparetti proposed an effective real time ranking algorithm using noise features [18].

2.3 Pseudo-Relevance Feedback via Query Expansion

Pseudo-relevance feedback techniques, represented as the relevance model [19], have been widely studied in information retrieval. Relevance model has been

improved by some researchers. Lv and Zhai proposed a model that optimizes the balance of the query and feedback information, and automatically learns the parameters of relevance model [20]. Tao and Zhai proposed a probabilistic mixture model using different parameters to each document and integrating the original query with feedback documents [21], and then this model was modified by Dillon and Collins-Thompson [22].

There have been work on term selection and document selection of relevance model. Cao et al. used SVM to classify good and bad terms [23]. Lv and Zhai extended relevance model to exploit term positions in the feedback documents [24]. Raman et al. chose terms that discriminate pseudo-relevant documents from pseudo-irrelevant documents [25]. Huang et al. proposed an approach to determine the optimal number of feedback documents with clarity score and cumulative gain [26]. He and Ounis used classification model to select good documents [27].

3 Motivation

3.1 Language Model in Microblog Retrieval

A statistical language model assigns a probability to a sequence of words by means of a probability distribution. In information retrieval, language model is used in the query likelihood model. Each document in the collection is represented as a language model. Documents are ranked based on the probability of query $Q = q_1, q_2, \ldots, q_n$ given document's language model $P(Q \mid M_D)$. Since authors usually use topic words repeatedly, keywords of document is expected to have large probabilities in the corresponding language model. The unigram language model is commonly used to achieve this.

$$P(Q \mid M_D) = \prod_{i=1}^{n} P(q_i \mid M_D) \tag{1}$$

$$P(q_i \mid M_D) = \frac{f_{q_i,D} + \mu \frac{c_{q_i}}{|C|}}{|D| + \mu} \tag{2}$$

When language model is used in microblog retrieval, one of the biggest challenge we have is that documents are too short. Most of the terms only appear once in one document. Keywords of document cannot be differentiated from other words in language model. Table 1 shows an example query from TREC Microblog Track 2011.

All the listed non-relevant documents and relevant documents contain two query terms. For example, relevant documents 2, non-relevant documents 1 and 2 all have "British" and "politician", but topic of non-relevant documents is not about the query. When we are calculating query likelihood as in Eq. (2), they all get 1 in $f_{q_i,D}$, and only difference will be at smoothing and document length. Therefore, language model does not work well in microblog retrieval.

Table 1. Examples of Retrieval Results

Query MB008: **phone hacking British politicians**
Non-Relevant Documents:
1. Boris Johnson has to be my favourite British politician of all time. He is an absolute LEGEND.
2. Politicians may be too nervous to address Britain's increasing irrelevance on the world stage, but they must
Relevant Documents:
1. British Tabloid Dismisses Editor Over Hacking Scandal
2. To Spy Politicians, British Aide to Prime Minister Resigns:
3. Ex-PM Brown feared voicemail hacking amid scandal: Former British Prime Minister Gordon Brown wrote to the police last summer to ask ...

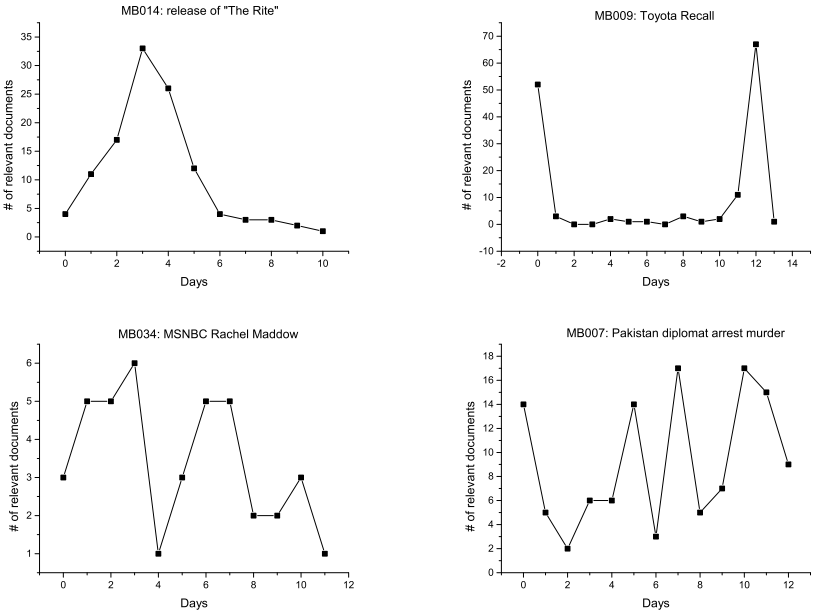

Fig. 1. Temporal in microblog retrieval

3.2 Temporal in Microblog Retrieval

Previous studies have shown that temporal distribution of relevant tweets is not uniform and should be considered in the ranking. Fig. 1 shows a visualization of four different types of query from TREC Microblog Track 2011. X-axis shows time prior to the query time, in days.

Query MB014 "release of 'The Rite' " has a single burst, which happens at the day of movie "The Rite" premiere. Query MB009 "Toyota Recall" has two

bursts, which are the day that Toyota initiated vehicle recalls and the day government announced the investigation report. Query MB007 "Pakistan diplomat arrest murder" has more than one burst. Although recency is an important fact, documents do not always cluster right before the query time.

4 Models

4.1 Feedback Language Model

The reason that language model does not perform well in microblog retrieval is that documents are short and most of query terms only show once. We need a better language model to describe topic of query. We adopted the idea of using document likelihood and relevance model [19]. Relevance model is a language model that represents the topic covered by relevant documents. Query Q can be seen as a small sample generated by the relevance model, and relevant document can be seen as a big sample generated by the same model. Document likelihood model use $P(D|R)$ as the probability of document generated by the given relevance model. In web page retrieval, document likelihood is very difficult to estimate since the variation of document length can be very large. One document may contains 10 terms while other may contains 10,000 terms. However, in microblog retrieval, we notice that lengths of tweets are not varied largely.

Here we use language model generated by the pseudo-relevance feedback documents to estimate document likelihood. Pseudo-relevance feedback documents $F = D_1, D_2, \ldots, D_k$ are the search results that returned from the first retrieval of the original query. Based on the idea of pseudo-relevance feedback, we assume that top k ranked documents are relevant. Feedback language model M_F is generated based on all the documents in F. Therefore, we can estimate probability of document generated by feedback language model.

$$P(D \mid M_F) = \prod_{w_i \in D} P(w_i \mid M_F) \tag{3}$$

$$P(w_i \mid M_F) = \frac{\sum_{D_j \in F} f_{w_i, D_j}}{\sum_{D_j \in F} |D_j|} \tag{4}$$

where f_{w_i,D_j} is the frequency of term w_i in document D_j.

Although lengths of tweets are not varied much, we apply normalization to Eq. (3) so that affects caused by different document length can be reduced [28].

$$P_{norm}(D \mid M_F) = APW^d \cdot P(D \mid M_F) \tag{5}$$

where APW^d denotes the penalty factor depending on document length. d equals to average document length subtracts length of D, and APW is average probability weight.

$$APW_{P(D|M_F)} = \frac{1}{|D|} \sum_{w_i \in D} P(w_i \mid M_F) \tag{6}$$

As described above, we generate document probability based on feedback language model. We now define the new score of document as a linear combination of scores produced by query likelihood model and feedback language model.

$$\begin{aligned} P'(D \mid Q) &= \lambda P(D \mid Q) + (1-\lambda)P(D \mid M_F) \\ &= \lambda \frac{P(Q \mid D)P(D)}{P(Q)} + (1-\lambda)P(D \mid M_F) \\ &= c\lambda P(Q \mid D) + (1-\lambda)P(D \mid M_F) \end{aligned} \tag{7}$$

where λ determines the weights of two models which is trained in the experiments, and constant $c = \frac{P(D)}{P(Q)}$ since $P(D)$ is usually assumed to be uniform.

4.2 Query Expansion Model

Query expansion is a well-studied technique to overcome the vocabulary mismatch problem in information retrieval. Several query expansion techniques have been developed. Pseudo-relevance feedback technique has been proven useful in previous work for improving retrieval performance.

Here we proposed a query expansion model that conducts both lexical expansions and temporal expansions. Our query expansion model is based on pseudo-relevance feedback technique. The idea of the proposed model is to expand original query with terms and times based on top-ranked documents from initial retrieval. For query $Q = q_1, q_2, \ldots, q_n$, we expand Q with:

1) Lexical expansion: expand original query with terms $Q_{lex} = w_1, w_2, \ldots, w_{lex}$.
2) Temporal expansion: expand original query with times $Q_{tem} = t_1, t_2, \ldots, t_{tem}$.

Here we have the new query $Q' = \{Q, Q_{lex}, Q_{tem}\}$.

We adopted framework proposed by Dakka et al. [5]. The framework assumed that document D can be split into a content component c_D and a temporal component t_D, and content relevance and temporal relevance are independent. The ranking function can be written as:

$$\begin{aligned} P(D \mid Q) &= P(c_D, t_D \mid Q) \\ &= P(c_D \mid Q)P(t_D \mid Q) \end{aligned} \tag{8}$$

c_D can been considered as D in language models.

Lexical Expansion. Relevance model generated expansion terms using pseudo-relevance feedback documents.

$$P(w \mid R) \propto \sum_{D \in R} P(w \mid D)P(D) \prod_{i=1}^{n} P(q_i \mid D) \tag{9}$$

Every terms from feedback documents are extracted and ranked according to Eq. (9). Top terms are chosen to expand the query. We interpolate the lexical expansion model with the retrieval model.

$$P(c_D \mid Q) \propto \alpha \sum_{w \in V} P(w \mid Q) \log P(w \mid D) + (1-\alpha) \sum_{w \in V} P(w \mid R) \log P(w \mid D) \tag{10}$$

Temporal Expansion. The idea of picking several times for temporal expansion is to build temporal profile for query and identify bursts in it. We take following steps to generate temporal expansion.

1. For query Q, get the top ranked documents $F_t = D_1, D_2, \ldots, D_t$.
2. Each document D has an associated time stamp, and we partition them into bins. Each bin corresponds to a time, for example days, hours, minutes. Number of bins depends on the time span of the document F_t.
3. Bins can be scored in two ways. The first way is to count the number of the documents in the bin. The second one is to add query likelihood scores of the documents in the bin.

$$score_{count}(bin(t)) = |D \in bin(t)| \tag{11}$$

$$score_{ql}(bin(t)) = \sum_{D \in bin(t)} P_{QL}(Q \mid D) \tag{12}$$

4. Rank bins based on their scores and expand the query using corresponding times of the top ranked bins.

After temporal expansions of query are generated, we can have temporal relevance $P(t_d \mid Q)$ as follows:

$$P(t_D \mid Q) = P(t_D \mid Q_{tem}) = P(t_D \mid t_1, t_2, \ldots, t_{tem}) \tag{13}$$

Given a serious of times, we use two ways to get the probability. The first one is to assume that the probability of document depends on the time that has the biggest impact of the document. The second one is to take the sum of all the impact of all the times. Thus we have:

$$P_{max}(t_D \mid t_1, t_2, \ldots, t_{tem}) = \max_{t_i \in Q_{tem}} P(t_D \mid t_i) \tag{14}$$

$$P_{sum}(t_D \mid t_1, t_2, \ldots, t_{tem}) = \sum_{t_i \in Q_{tem}} P(t_D \mid t_i) \tag{15}$$

To estimate the impact of temporal evidence, we use two ways to get $P(t_D \mid t_i)$.

1. Exponential function

$$P_{exp}(t_D \mid t_i) = e^{-\beta|t_i - t_D|} \tag{16}$$

2. Gaussian function

$$P_{gauss}(t_D \mid t_i) = e^{-\frac{|t_i - t_D|^2}{2\sigma^2}} \tag{17}$$

4.3 Feedback Model

We proposed our feedback model by combining our proposed feedback document model and query expansion model together. More specifically, we take the following steps:

1. Get initial retrieval results returned by original query.
2. Apply the proposed feedback language model to rerank the initial results.
3. Apply the proposed query expansion model to generate the new query with lexical expansions and temporal expansions.
4. Get retrieval results returned by the new query.
5. Apply the proposed feedback language model again to rerank the retrieval results.

5 Experiments

We have experimentally evaluated our model on TREC Microblog data. In Section 5.1, we first describe our experiment setup. Then we show the evaluation results of our proposed feedback language model in Section 5.2 and query expansion model in Section 5.3. In Section 5.4, we report the evaluation results of the feedback model, which is a combination of the feedback language model and the query expansion model. Finally, we conduct temporal query analysis by looking into different temporal types of queries in Section 5.5.

5.1 Setup

The experiments are conducted on TREC Microblog Track 2011 and 2012 data sets. The Track 2011 and 2012 evaluations are based on Track 2011 collection. The collection consists of an approximately 16 million tweets (1% sample of tweets from January 23, 2011 to February 7, 2011). There are 49 topics in Track 2011 and 59 topics in Track 2012 (MB050 topic and MB076 are deleted because of the absence of relevant documents). Each topic consists a query and its corresponding time stamp. Relevance judgements were based on a standard pooling strategy, and 3-point scale were used ("not relevant", "relevant", "highly relevant"). We removed all retweets and non-English tweets since TREC judged them as non-relevant. We indexed tweets posted before the time stamp associated with each topic using the Indri search engine[1]. No more than 1000 results are retrieved per topic.

[1] http://www.lemurproject.org/indri/

Table 2. Retrieval performance among non-expansion models

(a) TREC 2011

Methods	MAP	P@30	NDCG@30
QL	0.3082	0.3483	0.4254
SDM	0.2981	0.3463	0.4169
Recency Prior	$0.3112^{\dagger}$	0.3483	0.4330
FLM	$0.3202^{\dagger\ddagger}$	$0.3626^{\dagger\ddagger}$	$0.4417^{\dagger\ddagger}$

(b) TREC 2012

Methods	MAP	P@30	NDCG@30
QL	0.1868	0.2955	0.2836
SDM	0.1860	0.2955	0.2903
Recency Prior	0.1870	0.3006	0.2856
FLM	$0.1937^{\dagger\ddagger}$	0.3051	$0.2919^{\dagger}$

Each of our test models requires training data, we employ 2-fold cross-validation within each test collection. Parameters were trained with respect to precision at rank 30. We report mean average precision (MAP), precision at rank 30 (P@30) and NDCG at rank 30 (NDCG@30), which were the primary metrics used in the TREC Microblog evaluation. Statistical differences in our experiments are tested using a two-tailed paired t-test with level $\alpha = 5\%$.

5.2 Evaluation of Feedback Language Model

First we discuss the performance of our proposed feedback language model, referred to as **FLM**. Since this model doesn't involve query expansion, we picked several retrieval baselines without using query expansion techniques.

- **QL:** Standard query likelihood approach with Dirichlet smoothing ($\mu = 1500$).
- **SDM:** Sequential dependence model proposed by Metzler and Croft [29]. The model uses the original query words and bigrams extracted from the original query. We took default parameter settings, which are 0.85 for original query words, 0.15 for unwindowed bigrams, and 0.1 for windowed bigrams.
- **Recency Prior:** Recency prior for document is used in query likelihood model. It is one part of the time-based language model proposed by Li and Croft [4]. In $P(D|Q) \propto P(Q|D)P(D)$, $P(D)$ is assigned as a recency prior instead of being uniform. The recency prior is defined as $P(D) = \lambda e^{-\lambda(t_c - t_D)}$, where t_c is the query issued time and t_D is the time of the document.

Experiment results are shown in Table 2. Please note that † means performance of the method improves statistical significantly over QL baseline, and ‡ means performance of the method improve statistical significantly over both QL

Table 3. Retrieval performance on TREC 2011 among different variations of query expansion models

Methods	Description	MAP	P@30	NDCG@30
QEL	only lexical expansions	0.3217	0.3571	0.4431
QELX_ECM	exp + count + max	0.3396†	0.3823†	0.4643†
QELX_ECS	exp + count + sum	0.3348	0.3803	0.4595
QELX_EQM	exp + ql + max	0.3391†	0.3830†	0.4668†
QELX_EQS	exp + ql + sum	0.3360	0.3810	0.4645
QELX_GCM	gauss + count + max	0.3376†	0.3789†	0.4649†
QELX_GCS	gauss + count + sum	0.3367	0.3769	0.4644†
QELX_GQM	gauss + ql + max	0.3373†	0.3789	0.4646†
QELX_GQS	gauss + ql + sum	0.3373	0.3776	0.4653†

and *Recency Prior* baselines. Although *SDM* model has shown effectiveness in previous research of information retrieval, it fails in microblog retrieval. As we discussed above, documents are very short in microblog retrieval so that query bigrams are not likely to be seen in a certain window size. In TREC 2011, we can see that *Recency Prior* method helps the performance, but the improvements are not significant for all the metrics. Our proposed model outperforms both *QL* and *Recency Prior* baselines significantly. However, in TREC 2012, the performance of initial retrieval is not very effective so that feedback documents cannot provide much useful information. Although our proposed model has shown effectiveness on the performance, the improvements are not significant for all the metrics. *Recency Prior* method does not perform very well at this data because the temporal distribution of the query is not always clustered before the query time.

5.3 Evaluation of Query Expansion Model

In Section 4.2, we suggest different ways of getting three functions, which are $score(bin(t))$, $P(t_D \mid t_1, t_2, \ldots, t_{tem})$ and $P(t_D \mid t_i)$. To explore the effectiveness of our proposed different functions, we tested all the combinations of the functions. Experiment results are shown in Table 3. Descriptions of each abbreviation are also listed in the table. Please note that *QEL* denotes query expansion method with only lexical expansions, which is equivalent to the relevance model [19]. Notation † means performance of the method improves statistical significantly over *QEL*. Due to the limitation of space, we only demonstrate the results from TREC 2011.

We can see from the results that all eight combinations outperform query expansion method without temporal expansions. Some of them get significant improvements, while some of them do not. We think different combinations working with different types of queries. For example, methods with "sum" component work well with temporally unambiguous queries and methods with "max" component work well with temporally ambiguous queries. "exp" and "gauss"

Table 4. Retrieval performance among proposed feedback model and baseline models

(a) TREC 2011

Methods	MAP	P@30	NDCG@30
RM3	0.3261	0.3653	0.4504
Recency	0.3376	$0.3769^{\dagger}$	0.4632
QELX	$0.3391^{\dagger}$	$0.3830^{\dagger}$	$0.4668^{\dagger}$
MFM	$0.3520^{\dagger\dagger}$	$0.4027^{\dagger\ddagger}$	$0.4820^{\dagger\ddagger}$

(b) TREC 2012

Methods	MAP	P@30	NDCG@30
RM3	0.2006	0.3062	0.2902
Recency	0.2042	0.3051	0.2951
QELX	$0.2116^{\dagger\ddagger}$	$0.3232^{\dagger\ddagger}$	$0.3084^{\dagger\ddagger}$
MFM	$0.2122^{\dagger\ddagger}$	$0.3260^{\dagger\ddagger}$	$0.3107^{\dagger\ddagger}$

functions work similarly. We think more reasonable distance functions can be explored in the future. In the following experiments, we use *QELX* as a short for *QELX_EQM* to represent the query expansion model.

5.4 Evaluation of Feedback Model

We combine our proposed feedback document model and query expansion model together as the feedback model, referred to as **MFM**. To conduct the comparison experiments, we picked two retrieval baselines.

- **RM:** Relevance model proposed by Lavrenko and Croft [19].
- **Recency:** Time-based language models proposed by Li and Croft [4]. In *Recency Prior* method, recency prior is only used in the query likelihood model. Here, recency prior for document $P(D)$ is also used in relevance model $P(w \mid R) \propto \sum_{D \in R} P(w \mid D)P(D) \prod_{i=1}^{n} P(q_i \mid D)$.

We display the performance of the baselines and our proposed models in Table 4. Please note that † and ‡ mean the performance of the method is statistically significant over *RM* and *Recency* respectively. In TREC 2011, both of our proposed models *QELX* and *MFM* significantly outperform *RM3* baseline . The performance of *MFM* model is also significantly better than *Recency* baseline. We observe some improvements of *QELX* over *Recency* as well. In TREC 2012, the results show that our proposed models perform significantly better than both of the baselines.

5.5 Temporal Query Analysis

To explore how our proposed model performs on different types of temporal queries, we identify the number of bursts for each query using the relevant judgments. For each document D, we partition them into bins using their associated

Table 5. Examples of Retrieval Results

Track	#Bursts	#Topics	Topic
TREC 2011	0	1	33
	1	37	1-6, 10-18, 21, 22, 24, 27, 28, 30-33, 34-36, 38-45, 47-49
	2	9	8, 9, 19, 20, 23, 25, 26, 29, 37
	>2	2	7, 46
TREC 2012	1	33	52-54, 56, 57, 60, 62, 63, 65, 66, 68, 70-73, 75, 77, 80-82, 85, 86, 89, 91-94, 96, 99, 100, 104, 106, 108
	2	21	51, 58, 59, 61, 64, 67, 69, 74, 79, 83, 87, 88, 90, 95, 98, 101-103, 105, 109, 110
	>2	5	55, 78, 84, 97, 107

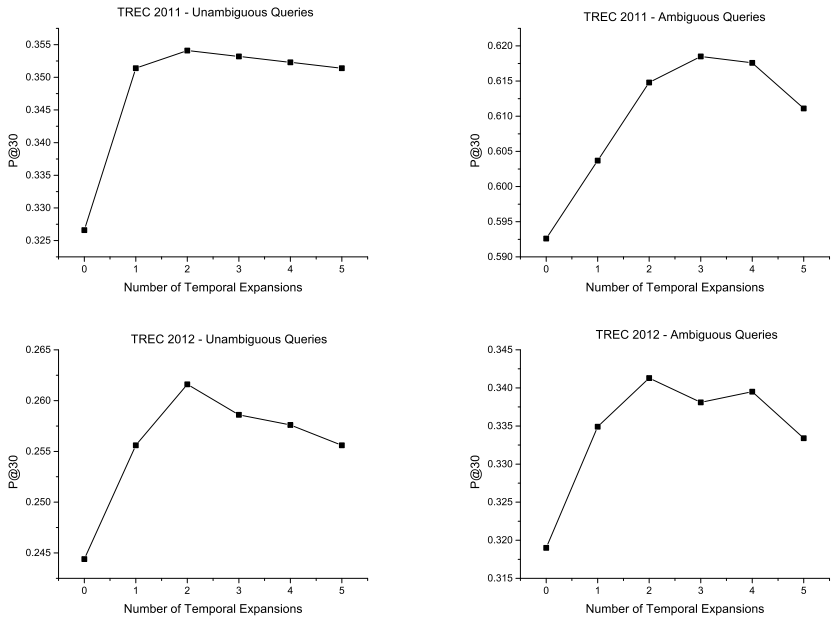

Fig. 2. Performance of queries with different temporal types with different numbers of bins

time stamps. Here "day" is used as time span of the bins. We use $B = \{b\}$ to represent set of bins. We define score of bins as $score(b) = |D \in b|$. Then we build the set of bursts U:

$$U = \{u \in B | \forall b \in (B - U), score(u) - score(b) >= \sigma\} \tag{18}$$

σ represents the standard variation of all the bins' scores.

Queries from TREC 2011 and 2012 classified by their number of bursts are listed in Table 5. We consider queries with zero or one burst as temporally unambiguous queries and queries with more than one bursts as temporally ambiguous

queries. For different types of queries, we tested how the performance changes according to different numbers of the bins. The experiment results are shown in Fig. 2. In TREC 2011, the performance of both query types get a big improvement by using the first temporal expansion. After that, the performance of temporally unambiguous queries begin to steady and then slightly decline, which is consistent with the ground truth that these queries only have one burst. For temporal ambiguous queries, the performance keeps growing and stops at the third expansion and then drops. In TREC 2012, the performance reaches the maximum value when adding two temporal expansions in both cases, then they slowly go down. From all the lines, we can see that when the number of temporal expansion are less than the number of bursts in the ground truth, adding more temporal expansions improves the performance. When the model using the same number of temporal expansion as the number of bursts, adding more temporal expansions also helps the performance in same cases. Overall, we can conclude that our proposed model involving temporal expansion improves the performance of not only temporally ambiguous queries but also temporally unambiguous queries.

6 Conclusion

In this paper, we proposed a feedback model for microblog retrieval. The feedback model includes a feedback language model and a query expansion model considering both lexical expansions and temporal expansions. Experiment results on TREC Microblog Track 2011 and 2012 data sets show that our proposed models improve upon existing baselines. Researchers have been paying growing attention to temporal evidence in information retrieval area. We think there is a lot more benefit that we can get from it. As future work, we will investigate more interesting and effective ways to use temporal information. It is also worth mentioning that there are many other directions in microblog retrieval that can be followed including using URLs information and modeling the noise and quality of tweets.

Acknowledgments. This paper is partially supported by the National Natural Science Foundation of China (NSFC Grant No. 61472006), the Doctoral Program of Higher Education of China (Grant No. 20130001110032) and the National Basic Research Program (973 Program No. 2014CB340405).

References

1. Efron, M., Golovchinsky, G.: Estimation methods for ranking recent information. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, pp. 495–504. ACM, New York (2011)
2. Keikha, M., Gerani, S., Crestani, F.: Time-based relevance models. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, pp. 1087–1088. ACM, New York (2011)

3. Liang, F., Qiang, R., Yang, J.: Exploiting real-time information retrieval in the microblogosphere. In: Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2012, pp. 267–276. ACM, New York (2012)
4. Li, X., Bruce Croft, W.: Time-based language models. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management, CIKM 2003, pp. 469–475. ACM, New York (2003)
5. Dakka, W., Gravano, L., Ipeirotis, P.G.: Answering general time sensitive queries. IEEE Transactions on Knowledge and Data Engineering **24**(2) (2012)
6. Massoudi, K., Tsagkias, M., de Rijke, M., Weerkamp, W.: Incorporating query expansion and quality indicators in searching microblog posts. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 362–367. Springer, Heidelberg (2011)
7. Jones, R., Diaz, F.: Temporal profiles of queries. ACM Trans. Inf. Syst. **25**(3), Article 14 (2007)
8. Liang, S., Ren, Z., Weerkamp, W., Meij, E., de, Rijke, M.: Time-Aware rank aggregation for microblog search. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management, CIKM 2014. ACM, Shanghai (2014)
9. 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 2011, pp. 1245–1246. ACM, New York (2011)
10. Whiting, S., Klampanos, I.A., Jose, J.M.: Temporal pseudo-relevance feedback in microblog retrieval. In: Baeza-Yates, R., de Vries, A.P., Zaragoza, H., Cambazoglu, B.B., Murdock, V., Lempel, R., Silvestri, F. (eds.) ECIR 2012. LNCS, vol. 7224, pp. 522–526. Springer, Heidelberg (2012)
11. Miyanishi, T., Seki, K., Uehara, K.: Time-aware latent concept expansion for microblog search. In: Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM, pp. 1–4. Ann Arbor, Michigan (2014)
12. Metzler, D., Cai, C., Hovy, E.: Structured event retrieval over microblog archives. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2012, pp. 646–655. Association for Computational Linguistics, Stroudsburg (2012)
13. Chen, L., Chun, L., Ziyu, L., Quan, Z.: Hybrid pseudo-relevance feedback for microblog retrieval. J. Inf. Sci. **39**(6), 773–788 (2013)
14. Bandyopadhyay, A., Ghosh, K., Majumder, P., Mitra, M.: Query expansion for microblog retrieval. IJWS **1**(4), 368–380 (2012)
15. Efron, M., Organisciak, P., Fenlon, K.: Improving retrieval of short texts through document expansion. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2012, pp. 911–920. ACM, New York (2012)
16. Han, Z., Li, X., Yang, M., Qi, H., Li, S., Zhao, T.: HIT at TREC 2012 microblog track. In: Proceedings of Text Retrieval Conference (2012)
17. Choi, J., Bruce Croft, W., Kim, J.Y: Quality models for microblog retrieval. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM 2012, pp. 1834–1838. ACM, New York (2012)
18. Gurini, D.F., Gasparetti, F.: Real-time algorithm for microblog ranking systems. In: Proceedings of The Twentyfirst Text Retrieval Conference, TREC 2012, Gaithersburg, pp. 6–9 (November 2012)

19. Lavrenko, V., Bruce Croft, W.: Relevance based language models. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2001, pp. 120–127. ACM, New York (2001)
20. Lv, Y., Zhai, C.X.: Adaptive relevance feedback in information retrieval. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 255–264. ACM, New York (2009)
21. Tao, T., Zhai, C.X.: Regularized estimation of mixture models for robust pseudo-relevance feedback. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2006, pp. 162–169. ACM, New York (2006)
22. Dillon, J.V., Collins-Thompson, K.: A unified optimization framework for robust pseudo-relevance feedback algorithms. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM 2010, pp. 1069–1078. ACM, New York (2010)
23. 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 on Research and Development in Information Retrieval, SIGIR 2008, pp. 243–250. ACM, New York (2008)
24. Lv, Y., Zhai, C.-X.: Positional relevance model for pseudo-relevance feedback. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, pp. 579–586. ACM, New York (2010)
25. Raman, K., Udupa, R., Bhattacharya, P., Bhole, A.: On improving pseudo-relevance feedback using pseudo-irrelevant documents. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 573–576. Springer, Heidelberg (2010)
26. Huang, Q., Song, D., Rüger, S.M.: Robust query-specific pseudo feedback document selection for query expansion. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 547–554. Springer, Heidelberg (2008)
27. He, B., Ounis, I.: Finding good feedback documents. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 2011–2014. ACM, New York (2009)
28. Maier, V.: Facing the problem of combining the language model with the acoustic model in speech recognition. Master Degree Thesis. University of Sheffield (2003)
29. Metzler, D., Croft, W.B.: A Markov random field model for term dependencies. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2005, pp. 472–479. ACM, New York (2005)