Improving Pseudo-Relevance Feedback via Tweet Selection

Taiki Miyanishi, Kazuhiro Seki, Kuniaki Uehara
Graduate School of System Informatics, Kobe University
{miyanishi, seki, uehara}@ai.cs.kobe-u.ac.jp

ABSTRACT

Query expansion methods using pseudo-relevance feedback have been shown effective for microblog search because they can solve vocabulary mismatch problems often seen in searching short documents such as Twitter messages (tweets), which are limited to 140 characters. Pseudo-relevance feedback assumes that the top ranked documents in the initial search results are relevant and that they contain topic-related words appropriate for relevance feedback. However, those assumptions do not always hold in reality because the initial search results often contain many irrelevant documents. In such a case, only a few of the suggested expansion words may be useful with many others being useless or even harmful. To overcome the limitation of pseudo-relevance feedback for microblog search, we propose a novel query expansion method based on two-stage relevance feedback that models search interests by manual tweet selection and integration of lexical and temporal evidence into its relevance model. Our experiments using a corpus of microblog data (the Tweets2011 corpus) demonstrate that the proposed two-stage relevance feedback approaches considerably improve search result relevance over almost all topics.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Query formulation, Relevance feedback

Keywords

Microblog search, Query expansion, Temporal dynamics

1. INTRODUCTION

Query expansion based on relevance feedback has been shown effective for improving microblog search performance [4, 19, 22, 23, 26]. That is due to the fact that query expansion can overcome the severe vocabulary mismatch problem of microblog search. However, classical relevance feedback, such as the Rocchio algorithm [31], requires a number of judged documents. Moreover, relevance judgment is often burdensome as it requires manually reading those documents. On the other hand, query expansion based on pseudo-relevance feedback (PRF) does not require judged documents [5, 14, 17, 18, 21, 24, 38]. The assumptions behind PRF are that the top ranked documents in the initial search results are relevant and that they include good words for query expansion. When the assumptions do not hold, PRF results in ineffective query expansion [25]—only a few of the suggested expansion words are useful and many others are either harmful or useless [3]. To overcome these problems, we propose a simple but effective query expansion method with manual selection of a single relevant document, which typically includes topic-related words. Using the selected document as query expansion words for a new query, we can re-retrieve more relevant documents and, based on the documents, estimate more accurate lexical and temporal evidence for improving the second-stage PRF described shortly. We designate this first-stage relevance feedback as *tweet selection feedback* for searching Twitter messages (i.e., tweets).

Previous works have also shown that time-based language modeling and relevance feedback approaches are effective for microblog search [4, 8, 9, 19, 22]. As described herein, we build on these findings and propose a novel PRF method combining lexical and document-dependent temporal evidence of microblog in response to a query, which relies strongly on relevance information among the re-retrieved documents, such as a word distribution and a time-stamp distribution. We assume that the proposed PRF method further improves microblog search performance in combination with tweet selection feedback. To demonstrate the validity of our proposed approach, we carry out evaluative experiments on the datasets of the TREC 2011 and 2012 real-time ad-hoc task (i.e., Tweets2011 corpus[1]), which consist of more than 16 million tweets over a period of two weeks. The experimental results of the two-stage relevance feedback show that our tweet selection feedback reduces the adverse effects of PRF for difficult queries and is especially effective when combined with our proposed PRF.

The paper is structured as follows: In Section 2 we introduce the established pseudo-relevance feedback approaches. In Section 3 we present the limitation of standard relevance feedback methods. Section 4 describes details of our proposed method, which consists of two-stage relevance feedback, tweet selection feedback and lexical-and-temporal-based relevance feedback. In Section 5 we demonstrate the effect

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.2505701.

[1] http://trec.nist.gov/data/tweets/

of the proposed PRF methods. In Section 6 we survey related work. Finally, Section 7 presents a summary of this work and conclusions.

2. PREVIOUS METHODS OF TIME-BASED RELEVANCE FEEDBACK

2.1 Time-based Language Model for Retrieval

Query likelihood model. Our PRF model builds on language modeling frameworks for information retrieval (IR), particularly the query likelihood model as proposed by Ponte and Croft [30]. This model assumes the probability of a query Q as being generated by the word probabilities on a document D. Based on the language modeling approach, all documents are ranked in order of the posterior probability, which is defined as $P(D|Q)$. The probability of a document $P(D|Q)$ by Bayes' rule becomes

$$P(D|Q) \propto P(Q|D)P(D), \tag{1}$$

where $P(Q|D)$ is the query likelihood on the given document and $P(D)$ is the prior probability that D is relevant to any query. To capture word frequency information in indexing a document, the multinomial model is used. This is called a uni-gram language model. We have the query likelihood $P(Q|D)$ as follows:

$$P(Q|D) = \prod_{i=1}^{|Q|} P(w_i|D), \tag{2}$$

where $|Q|$ is the number of words in the query and $P(w|D)$ is the probability of a word w under the word distribution for a document D. 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 [39] is often used.

$$P(w|D) = \frac{|D|}{|D|+\mu} P_{ml}(w|D) + \frac{\mu}{|D|+\mu} P(w|C) \tag{3}$$

Therein, $P_{ml}(w|D) = \frac{c(w;D)}{\sum_{w' \in V} c(w';D)}$, $c(w;D)$ denotes the number of word counts of w in a document D, $P(w|C)$ is the collection language model. μ is the Dirichlet prior.

Recency-based language model. If we assume that the prior probability distribution over documents is uniform, then we rank documents in decreasing order of the query likelihood $P(Q|D)$ above. However, the quality of document is changing over time. Topically relevant but obsolete documents might not satisfy the user if recent information is preferred. Consequently, Li and Croft [18] incorporated a prior distribution considering recency over documents into language model frameworks for retrieval. They proposed application of the following exponential distribution as the document prior $P(D)$ to Eq. 1. We have

$$P(D|t_D) = r \cdot e^{-r \cdot |t_Q - t_D|}, \tag{4}$$

where t_Q stands for the query time at which a query was issued by a user, t_D signifies a time-stamp of the document D, and r denotes a rate parameter of the exponential distribution. This model includes the assumption that newer documents have a higher probability than older ones do.

2.2 Temporal Relevance Model

Pseudo-relevance model. Lavrenko and Croft [17] incorporated relevance feedback into language modeling frameworks. They estimated a relevance model, $P(w|R)$, using a joint probability of observing the word w together with query words on top ranked initial search results. That relevance model weights words w according to the following.

$$P(w|R) \approx P(w|Q) \propto \sum_{D_i \in \mathcal{R}} P(D_i) P(w|D_i) \prod_{j}^{|Q|} P(w_j|D_i) \tag{5}$$

Among those expressions, $\mathcal{R}$ is the top M retrieved documents using the query Q. This approach is called pseudo-relevance feedback. In addition, for query expansion, words w were ordered in descending order. The top K words are added to the original user query.

Recency-based relevance model. In addition, Li and Croft [18] incorporated recency into the relevance model redesigning the document prior as follows:

$$P(w|Q) \propto \sum_{D_i \in \mathcal{R}} P(D|t_D) P(w|D_i) \prod_{j}^{|Q|} P(w_j|D_i), \tag{6}$$

where $P(D|t_D)$ denotes the recency-based document prior in Eq. 4. This model is good at dealing with recency queries, but it is not able to accommodate any temporal variation. On microblog services, temporal dynamics of the topic varies, so that the recency-based method fails to find topic-related words having specific temporal variations consisting of an old peak that is distant from the query-time or a multimodal temporal variation [26]. Furthermore, this model was not able to accommodate query-specific recency even though the degree of recency is topic-dependent [7, 8].

Time-based relevance model. Keikha et al. [14] proposed a time-based relevance model. They assume that any topic relates to specific time and that their topic-related words are frequently used in this time. Their approach detects this topic-related time and incorporates this temporal property into language modeling frameworks as

$$P(w|Q) = \sum_{t} P(w|t,Q) P(t|Q). \tag{7}$$

The previous version by Choi and Croft [4] defined the word distribution $P(w|t,Q)$ at time t against a query Q as

$$P(w|t,Q) = \sum_{D_i \in \mathcal{R}_t} P(w|D_i) \prod_{j}^{|Q|} P(w_j|D_i), \tag{8}$$

where $\mathcal{R}_t$ represents the top M documents issued in time t. Although the original work by Keikha et al. [14] assumed $P(w|t,Q)$ was uniform, Choi and Croft assumed that $P(w|t,Q)$ was equal to $P(w|Q)$ and incorporated the time property into only $P(t|Q)$. This equation is the same to Eq. 5 when using documents in time t except for $P(D)$ is set to be uniform, so that their model can consider word probability information in time t. Consequently, Eq. 7 is interpreted as the weighted sum of $P(w|t,Q)$ by a temporal model $P(t|Q)$. The temporal model against a given query, $P(t|Q)$, is defined as

$$P(t|Q) = \frac{1}{Z} \sum_{D \in \mathcal{R}} P(t|D) P(Q|D), \tag{9}$$

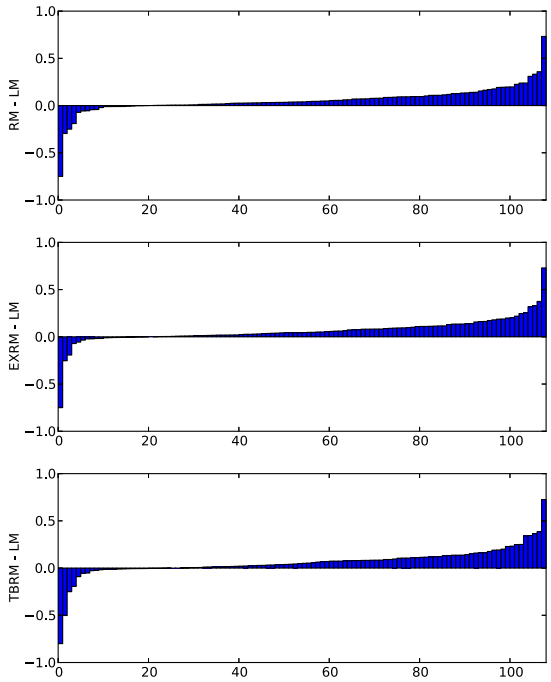

Figure 1: Improvements by existing relevance feedback methods over the initial search. Each bar shows the difference in average precision comparing LM to RM (top), EXRM (middle), and TBRM (bottom).

where $P(t|D)$ is an indicator function $P(t|D) = 1$ if the date of t and a document time-stamp of D is the same; otherwise, $P(t|D) = 0$. One must recall that $P(Q|D)$ is the query likelihood of a document D for Q. Z is the normalization factor. It is particularly interesting that this definition is the same as the notion of the temporal profile proposed by Jones and Diaz [12]. This model estimates topic-related time using document time-stamps and search scores (i.e. query likelihoods assuming the prior probability of document $P(D)$ is uniform) of retrieved documents. This relevance model can weight the word distribution by this temporal profile, so it is able to capture general temporal variation by each topic. However, it ignores recency and document-dependent temporal information.

3. LIMITATION OF PSEUDO-RELEVANCE FEEDBACK

The previous time-based language models for IR and temporal relevance model based on PRF integrated into query expansion methods achieved great success for improving microblog search performance [4, 8, 9, 19, 22]. They can incorporate recency or temporal variation on microblogging platform into their model and overcome the vocabulary mismatch problem. These PRF methods assume that the proportion of relevant documents in initial search results is large, so that top ranked documents include good words for query expansion. However, that assumption becomes invalid and PRF fails if the initial search rank non-relevant documents at the top [25]. Moreover, several words suggested by PRF model are useful and many others are either harmful or useless [3]. We assume that PRF for microblog search also

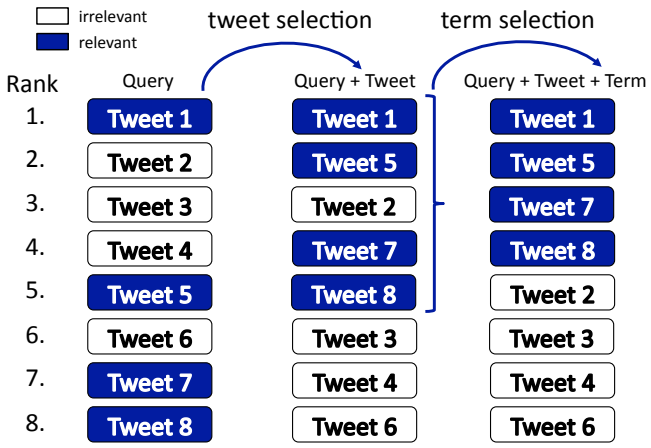

Figure 2: Overview of two-stage relevance feedback.

fails to improve search performance for some topics while enhancing the performance for other topics.

To see the performance of PRF over initial search results, we compare several PRF methods to the initial search. As the initial search, we use the language model with Dirichlet smoothing of Indri search engine[2]. We refer to this method as LM. Unless otherwise specified, all retrievals are implemented on top of LM. We prepare three baseline PRF methods: the standard relevance model [17] (see Eq. 5), exponential recency-based relevance model [18] (see Eq. 6), and time-based relevance model [4, 14] (see Eq. 7), which are respectively designated as RM, EXRM, and TBRM. The parameters of these PRF models are tuned. The parameter tuning and pre-process are discussed in Section 5.1 and 5.2. Figure 1 shows the bar plots of the difference in average precision of existing relevance models (RM, EXRM, and TBRM) over initial search results (LM) using 108 search topics for TREC 2011 and 2012 microblog track. Results showed that all PRF methods improved search performance for many topics, but simultaneously they decrease for several topics. The results imply that we must estimate more accurate temporal and lexical evidence for maintaining PRF performance and to improve microblog retrieval simultaneously.

4. PROPOSED METHOD

To overcome the limitation of established PRF methods and to improve retrieval further, we propose two-stage relevance feedback methods. They consist of tweet selection feedback (TSF) and query-document dependent temporal relevance model. We describe an overview of our approach in Figure 2. For the former, we only select a single relevant tweet among initial search results and re-retrieve tweets using the selected tweet as expansion words of a new query. For the latter, we apply query-document dependent temporal query expansion method to the re-retrieved documents, which almost all include relevant tweets at the top. The following sections show details of the respective methods.

4.1 Tweet Selection Feedback

The first relevance feedback uses a selected tweet from the initial search results. We assume that the relevant tweet se-

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

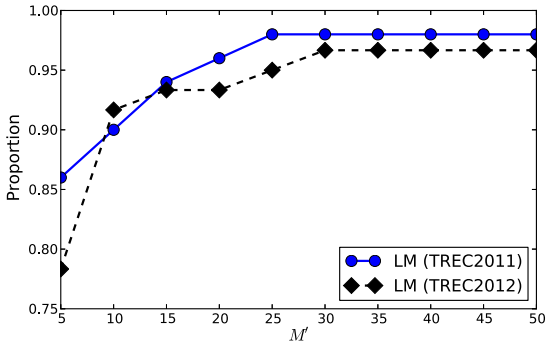

Figure 3: Proportion that at least one relevant document is contained among initial search results across different values of the cut off parameter M'.

lected by users is a good indicator to retrieve relevant tweets to a given query because the relevant tweet generally includes good topic-related words. Using the selected tweet as expansion word for re-retrieving documents, we can obtain relevant tweets similar to the selected tweet at the top.

Additionally, we observed that the top ranked tweets retrieved at the top by a standard search engine with default settings (LM in our case) are often relevant, so that users can easily detect at least a relevant document from top ranked documents. To see the initial search performance, we define the proportion of search topics that retrieve at least a single relevant document among the top M' documents. We have

$$\frac{1}{N'}\sum_{i=1}^{N'}\psi(P_i@M') \tag{10}$$

where $\psi(\cdot)$ is a function $\psi(x)=1$ if $x>0$; otherwise, $\psi(x)=0$. N' is the number of topics used and $P_i@M'$ is the value of precision at M' for the i-th topic. Figure 3 presents the proportion across several cut off parameters M' using TREC 2011 and 2012 microblog track topics. Results show that users can find relevant tweets at the top without much effort in the case of many TREC search topics. For example, the proportion of finding at least one relevant document among the top 30 is more than 0.95 in both datasets. Furthermore, users can read many tweets quickly because the length of the tweet content is limited to 140 characters. Consequently, users can readily detect a relevant tweet without much effort.

4.2 Query-Document Dependent Temporal Relevance Model

In this section, we introduce the query-document dependent temporal relevance model. We assume that the search results by tweet selection feedback rank many relevant documents at the top, which contains more accurate word and temporal distributions than by initial search. To use the improved pseudo-relevance information effectively, we propose a novel relevance feedback approach using lexical and temporal evidence.

We rely mainly on the notion of Dakka et al. [5] and Efron and Golovchinsky [8] for time-sensitive language modeling frameworks and also use a document expansion approach proposed by Efron et al. [9] to capture document-dependent temporal variation. We explain the relevance model step-by-step. First, we decompose a document part D in $P(w|Q)$ into the lexical word in document D_w and temporal information of document D_t following Dakka et al. [5],

$$\begin{aligned} P(w|Q) &= \sum_{D\in\mathcal{R}} P(w,D|Q) \\ &= \sum_{D\in\mathcal{R}} P(w,D_w,D_t|Q) \\ &= \sum_{D\in\mathcal{R}} P(w,D_w|D_t,Q)P(D_t|Q). \end{aligned} \tag{11}$$

Then, following Efron and Golovchinsky [8]'s work, we applied the simple assumption that the temporal relevance of D_t is independent of the document's content, D_w, and drop D_t from the conditional probability in Eq. 11. Moreover, we assume that the given query consisting of query words in Q and the words w in pseudo-relevant documents are sampled identically and independently from a uni-gram distribution of $\mathcal{R}$. Therefore, we have

$$\begin{aligned} P(w|Q) &= \sum_{D\in\mathcal{R}} P(w,D_w|Q)P(D_t|Q) \\ &\propto \sum_{D\in\mathcal{R}} \underbrace{P(w|D_w)P(Q|D_w)P(D_w)}_{\textbf{Lexical}}\underbrace{P(D_t|Q)}_{\textbf{Temporal}} \\ &= \sum_{D\in\mathcal{R}} P(w|D_w)\prod_{i}^{|Q|}P(q_i|D_w)P(D_t|Q), \end{aligned} \tag{12}$$

where $P(D_t|Q)$ is the query-dependent document generation probability from a temporal perspective. We designate $P(D_t|Q)$ as *temporal evidence*. However, $P(w|D_w)P(Q|D_w)P(D_w)$ is equal to a factor of the standard relevance feedback model (see Eq. 5). We designate this as *lexical evidence*. In Eq. 12, we assume that the prior probability over documents from a lexical perspective, $P(D_w)$, is uniform. Eq. 12 is the weighted sum of query-dependent lexical evidence by query-dependent temporal evidence with respect to each document.

Ideally, the probability $P(D_t|Q)$ becomes high when the query Q and the document D share a similar temporal property, so that we quantify this temporal property as the distance between two temporal models of Q and D using the notion of temporal profile [12]. Borrowing the idea of temporal profile in Eq. 9, we define the temporal models of Q and D as $P(t|Q)$ and $P(t|Q_D)$, respectively, where Q_D is the pseudo-query of D submitted to search engines as a query based on the idea of Efron et al. [9]. Using $P(t|Q_D)$, we can capture document-dependent temporal variation. In addition, we apply background smoothing to both temporal models and then smooth them with the model for adjacent days following previous works [5, 12]. Additionally, we assume that the distance between two temporal models approximately follows an exponential distribution because the documents retrieved by Q_D share more similar temporal property with the documents retrieved by Q than unobserved documents. We define the probability of query-dependent document's temporal evidence as

$$P(D_t|Q)\propto P(X>d)=e^{-\gamma d}, \tag{13}$$

where d is the distance of two temporal models between $P(t|Q)$ and $P(t|Q_D)$ and γ is a rate parameter of exponential distribution. Moreover, past works [7, 8] have shown that incorporating query-dependent recency is effective for improving microblog search. Therefore, we design the rate

parameter as automatically changing in response to each query's temporal property, as

$$\gamma = 1 - \sum_{t \in \mathcal{T}_Q} P(t|Q). \quad (14)$$

where $\mathcal{T}_Q = \{t \in \mathcal{T} : t_Q - t < \alpha\}$, $\mathcal{T}$ is a time range in a collection (days in our case), t_Q denotes a query-time of query Q, and α is a hyper-parameter that controls the impact of topic-recency. The probability γ denotes the value of complementary cumulative distribution function of temporal model until α days before the topic's query-time. If the temporal profile of a given query ranks many documents generated at around its query-time at the top, then the probability γ is low. However, the probability is high if those document time-stamps are far from the query-time.

We assume that similar temporal models should share similar temporal property (e.g. temporal variation). Therefore, we compare two temporal models using the Bhattacharyya coefficient,

$$\mathcal{B}(Q, D) = \sum_{t \in \mathcal{T}} \sqrt{P(t|Q)P(t|Q_D)}. \quad (15)$$

This comparison provides a similarity score between 0 and 1. Similar methods have been used to compare two associated language models using the Bhattacharyya coefficient [6]. Using the Bhattacharyya coefficient, we can obtain the distance between two temporal models as

$$d = -\ln \mathcal{B}(Q, D). \quad (16)$$

This is called Bhattacharyya distance. When we substitute Eq. 16 into Eq. 13, we have the following final equation:

$$P(D_t|Q) \propto \{\mathcal{B}(Q, D)\}^{\gamma}. \quad (17)$$

This probability $P(D_t|Q)$ becomes high when $P(t|Q)$ and $P(t|Q_D)$ are similar (i.e. Bhattacharyya coefficient is high). The increase of $P(D_t|Q)$ approaches linear increase when γ is high; in other words, a given topic indicates an old event. However, $P(D_t|Q)$ rapidly increases when a given topic indicates a recent event (i.e. γ is low).

5. EVALUATION

5.1 Experimental Setup

We evaluate our proposed methods using the test collection for the TREC 2011 and 2012 microblog track (Tweets2011 corpus). This collection consists of about 16 million tweets sampled between January 23 and February 8, 2011. In addition, 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 methods as *allrel* and *highrel*, where allrel considers both minimally relevant and highly relevant tweets as relevant and highrel considers only highly relevant tweets as relevant.

We indexed tweets posted before the specific time associated with each topic by the Indri search engine with the following setting. All queries and tweets are stemmed using the Krovetz stemmer without stop-word removal. They are case-insensitive. This index was created to simulate a realistic real-time search setting, where no future information is available when a query is issued. We built an index for each query. In our experiments, we used the titles of TREC topics numbered 1–50 and 51–110[3] as test queries, which are the official queries in the TREC 2011 and 2012 microblog track, respectively. Additionally, we used 33 topics at TREC 2011 and 56 topics at TREC 2012, and obtained highly relevant tweets for highrel.

For retrieving documents, we used a basic query likelihood model with Dirichlet smoothing [39] (we set smoothing parameter $\mu = 2500$ similar to Efron's work [9]) implemented by the Indri search engine [34] as the language model for IR (LM) and all PRF used this LM as initial search results. We filtered out all non-English retrieved tweets using a language detector with infinity-gram, called ldig[4]. The retweets[5] were regarded as irrelevant for evaluation in the TREC microblog track [27, 33]; however, we used retweets except in a final ranking of tweets because a set of retweets is a good source might contain topic-related words for improving Twitter search performance [4]. In accordance with the track's guidelines, all tweets with http status codes of 301, 302, 403, and 404 and all retweets contain the string "RT" at the beginning of the tweet were removed from the final ranking. Finally, we used the top 1000 results for evaluation.

5.2 Baselines

Our approach first conducts tweet selection feedback (TSF) described in Section 4.1. We automatically select relevant tweets from initial search results among top L tweets for TSF by each topic. We set L to 30 based on a preliminary experiment. In Section 5.4, we show that the performance is not sensitive to the choice of L when L is sufficiently large (e.g. $L \geq 30$). The selected relevant tweets are minimally or highly relevant tweets. When multiple relevant tweets exist in initial search results, we use only a single relevant tweet that contains more words in it than others. We assume that users prefer long tweets. 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 Indri's 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 based on a query-and-document dependent temporal relevance model (QDRM). For QDRM, we produce a temporal profile consisting of the top N tweets, which were retrieved using a document among initial search results as a pseudo-query. These pseudo-queries were also pre-processed in the same mode as tweets used for TSF. We denote the combination of TSF and QDRM as TSF + QDRM.

To assess our proposed methods, TSF and TSF + QDRM, we also prepared several baseline methods. Our first baseline, RM, uses standard relevance feedback using only lexical evidence [17]. This can be compared with TSF + RM which uses tweet selection feedback before the pseudo-relevance feedback RM. QDRM differs from RM in that RM does not consider temporal evidence. Actually, QDRM is equal to RM when we set γ in QDRM to 0 (see Eqs. 12 and 13). Our second baseline, EXRM uses relevance feedback using exponen-

[3]The topic numbered MB050 and MB076 has no minimally or highly relevant tweets. Therefore, we did not use them for our experiments.

[4]https://github.com/shuyo/ldig

[5]Tweets re-posted by another user to share information with other users

Table 1: Performance comparison of the proposed methods and baselines for allrel documents.

	TREC 2011				TREC 2012			
Method	AP	nDCG@10	P@10	P@30	AP	nDCG@10	P@10	P@30
LM	0.3571	0.5301	0.4755	0.4143	0.2408	0.4177	0.4814	0.3847
RM	0.4063_{l}	0.5616	0.5673_{l}	0.4741_{l}	0.3024_{l}	0.4592_{l}	0.5475_{l}	0.4503_{l}
EXRM	0.4204_{lr}	0.5725	0.5816_{l}	0.4762_{l}	0.3025_{l}	0.4663_{l}	0.5492_{l}	0.4520_{l}
TBRM	0.4020	0.5573	0.5673_{l}	0.4728_{l}	0.3139_{l}	0.4826_{l}	0.5610_{l}	0.4644_{lq}
QDRM	0.4206_{l}	0.5843	0.5735_{l}	0.4721_{l}	0.3039_{l}	0.4760_{l}	0.5542_{l}	0.4441_{l}
TSF + LM	$0.5040^{\blacktriangle}$	$\mathbf{0.6956}^{\blacktriangle}$	$0.6388^{\blacktriangle}$	$0.4966^{\blacktriangle}$	$0.3198^{\blacktriangle}$	$0.5309^{\blacktriangle}$	$0.5763^{\blacktriangle}$	$0.4559^{\blacktriangle}$
TSF + RM	$0.5287^{\blacktriangle}$	$0.6730^{\blacktriangle}$	$0.6327^{\triangle}$	$0.5224_{l'}$	$0.3475^{\triangle}_{l'}$	$0.5352^{\triangle}$	$0.6068^{\triangle}$	0.4785
TSF + EXRM	$0.5328^{\blacktriangle}$	$0.6814^{\blacktriangle}$	$0.6449^{\triangle}$	$0.5218^{\triangle}_{l'}$	$0.3476^{\triangle}_{l't'}$	0.5329	$0.6068^{\triangle}$	0.4797
TSF + TBRM	$0.5174^{\blacktriangle}$	$0.6745^{\blacktriangle}$	$0.6429^{\blacktriangle}$	0.5177	$0.3415_{l'}$	0.5331	0.6051	0.4763
TSF + QDRM	$\mathbf{0.5384}^{\blacktriangle}_{l'}$	$0.6843^{\blacktriangle}$	$\mathbf{0.6571}^{\blacktriangle}_{r'}$	$\mathbf{0.5354}^{\blacktriangle}_{l'}$	$\mathbf{0.3584}^{\blacktriangle}_{l'r'e't'}$	$\mathbf{0.5552}^{\blacktriangle}_{r'e'}$	$\mathbf{0.6220}^{\blacktriangle}$	$\mathbf{0.4910}^{\blacktriangle}_{l'}$

tial distribution to prior probability for relevance model [18]. EXRM does not consider query-dependent recency and temporal variation compared to QDRM. We also prepare TSF + EXRM, which is a combination of TSF and EXRM to assess the effect of tweet selection feedback for the recency-based method. Finally, our third baseline is a time-based relevance model, TBRM, that incorporates lexical evidence and query-dependent temporal variation into its relevance model. However, it ignores recency and document-dependent temporal variation. We compare this model and its tweet selection extension, TSF + TBRM, to our QDRM that uses both lexical and temporal evidence with query-dependent recency. RM, EXRM, and TBRM are strong baselines in our experiments.

For all query expansion methods, we select candidate words among the top M tweets retrieved by the original query after removing the uniform resource locators (URLs), and user names starting with '@' or special characters (!, @, #, ', ", etc.). All query words, candidate words, and tweets are decapitalized. The candidate words include no stopwords prepared in the Indri search engine. Then, we select K words among candidate words in descending order of the probability $P(w|Q)$, respectively. The selected words contain no original query word, but might contain words of the selected tweet in the case of using TSF. Finally, we combined the expanded words of PRF and the original query (or the combination of the original query and the selected tweet) as an expanded query; they were weighted with $1:1$.

For QDRM and EXRM, we tune parameters: the length of temporal profile (i.e. N), the hyper-parameter (i.e. α), and the rate parameter (i.e. r). For all methods, we also tune their parameters: the number of pseudo-relevance feedback documents (i.e. M) and the number of expansion words (i.e. K). Values of the model parameters are optimized for best performance precision at 30 on training data, which is the official measure in TREC 2011 microblog track. For example, we tune parameters of the IR model using TREC 2012 microblog track dataset and test it with TREC 2011 microblog dataset. However, we trained the model using the TREC 2012 dataset and test it on the TREC 2011 dataset. Results show that the parameter N in the proposed QDRM set to be 10 is better for both datasets. The sensitivity of other important parameters such as L in TSF and the recency control parameter α of QDRM is discussed in the next section.

5.3 Evaluation Measure

The goal of our system is to return a ranked list of tweets using (pseudo-) relevance feedback methods. To evaluate retrieval effectiveness, we used precision at 10 and 30 (P@10, P@30, respectively), average precision (AP), and normalized discounted cumulative gain (nDCG) [11], nDCG considers graded relevance. Recall that P@30 was the official Microblog track metric in 2011. In the TREC 2012 microblog track, "highly relevant" tweets are the required level of relevance. These measures provide a succinct summary of the quality of the retrieved tweets. We discuss the statistical significance of results obtained using a permutation test [32] throughout this paper.

5.4 Experimental Results

Overall Results. Table 1 shows the P@10, P@30, AP, and nDCG performances of 10 methods with statistical significance test results for allrel documents. Table 2 shows the P@30 and AP performances for highly relevant documents. Significant improvements by tweet selection feedback (TSF) are denoted with $\triangle$ and $\blacktriangle$, respectively, for significance probabilities $p < 0.05$ and $p < 0.01$. In addition, among methods without the use of TSF, the subscript l, r, e, t, and q respectively indicate statistically significant improvements ($p < 0.05$) over LM, RM, EXRM, TBRM, and QDRM. Moreover, among methods using TSF, the subscripts l', r', e', t', and q' respectively indicate statistically significant improvements ($p < 0.05$) over TSF + LM, TSF + RM, TSF + EXRM, TSF + TBRM, and TSF + QDRM. The best result per column is marked in bold typeface.

It is apparent that QDRM markedly outperforms the initial search LM on most measures across both datasets, similarly to other relevance feedback approaches RM, EXRM, and TBRM with statistical significance. Moreover, QDRM outperformed the standard relevance model RM in terms of most evaluation measures across both datasets similar to other time-based relevance feedback methods EXRM and TBRM, which suggests that temporal evidence (recency or temporal variation) is important for microblog search. However, none of these differences is statistically significant except between RM and EXRM on AP.

When using tweet selection feedback, TSF + LM markedly outperformed LM in terms of all measures across both datasets with statistical significance, which suggests that the simple query expansion method using a selected relevant tweet as expansion words is considerably effective. Furthermore, relevance feedback approaches after TSF outperformed relevance feedback without using TSF in terms of all measures. For all using TSF, the differences in AP, nDCG@10, and P@10 in the TREC 2011 dataset were statistically signifi-

Table 2: Performance comparison of the proposed method and baselines for highrel documents of TREC 2011 and 2012 datasets.

	TREC 2011		TREC 2012	
Method	AP	P@30	AP	P@30
LM	0.2747	0.1293	0.1766	0.1976
RM	0.2499	0.1374	0.2258_{l}	0.2494_{l}
EXRM	0.2710_{t}	0.1465_{r}	0.2270_{l}	0.2548_{l}
TBRM	0.2404	0.1374	0.2314_{l}	0.2583_{l}
QDRM	0.2911	0.1424	0.2293_{l}	0.2500_{l}
TSF + LM	0.3461^{Δ}	0.1566	$0.2180^{\blacktriangle}$	$0.2387^{\blacktriangle}$
TSF + RM	0.3508^{Δ}	0.1727^{Δ}	0.2358	0.2595
TSF + EXRM	0.3476	0.1747	0.2358	0.2613
TSF + TBRM	0.3365^{Δ}	0.1717	0.2325	0.2542
TSF + QDRM	$\mathbf{0.3619}_{l'}$	$\mathbf{0.1758}^{\Delta}$	$\mathbf{0.2389}_{l'}$	$\mathbf{0.2649}_{l'}$

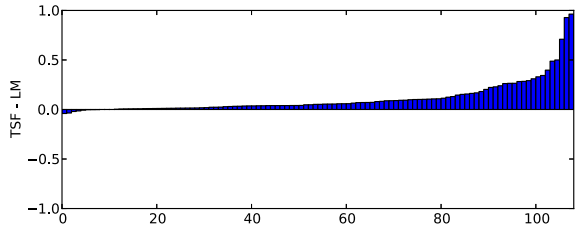

Figure 4: Difference in average precision between TSF and LM using the TREC 2011 and 2012 microblog track topics.

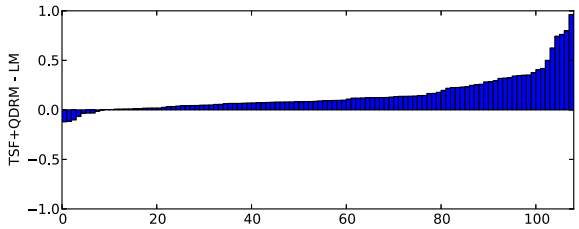

Figure 5: Difference in average precision between TSF + QDRM and LM using the TREC 2011 and 2012 microblog track topics.

cant. Important points include the fact that TSF + QDRM markedly outperformed QDRM with regard to all evaluation measures across both datasets with statistical significance. For both datasets, TSF + QDRM outperformed other PRF methods using TSF: TSF + RM, TSF + EXRM, and TSF + TBRM. Particularly the difference in average precision on the TREC 2012 dataset is statistically significant. Results suggest that tweet selection feedback is useful for PRF methods and that incorporating query-dependent lexical and temporal evidence by each document is considerably effective when using improved search results by tweet selection feedback.

From Table 2, it is also apparent that PRF using TSF is effective for improving retrieval performance when searching highly relevant documents. In this case, TSF + QDRM outperformed other methods in all evaluation measures across both datasets. For further improvement of search performance with regard to highly relevant documents, we must consider external web-contents corresponding to URLs in a tweet, which significantly affect the retrieval performance of highly relevant tweets [19].

Table 3: Improved and decreased percentages of the values of mean average precision (MAP [%]) and the number of topics (#) by pseudo-relevance feedback methods over the initial search using the TREC 2011 and 2012 topics.

	Improved		Decreased	
Method	MAP [%]	#	MAP [%]	#
RM	51.5 (93.9)	87	-37.0 (24.8)	20
EXRM	55.4 (96.2)	87	-30.5 (26.6)	20
TBRM	64.4 (107.2)	81	-29.5 (24.9)	25
QDRM	46.3 (65.1)	86	-21.8 (18.2)	18
TSF + LM	122.4 (314.9)	97	-4.8 (4.4)	8
TSF + RM	161.4 (350.6)	92	-25.4 (17.0)	14
TSF + EXRM	162.2 (348.2)	92	-25.5 (17.0)	14
TSF + TBRM	160.6 (350.9)	91	-28.7 (20.1)	16
TSF + QDRM	153.7 (337.6)	97	-26.3 (19.0)	9

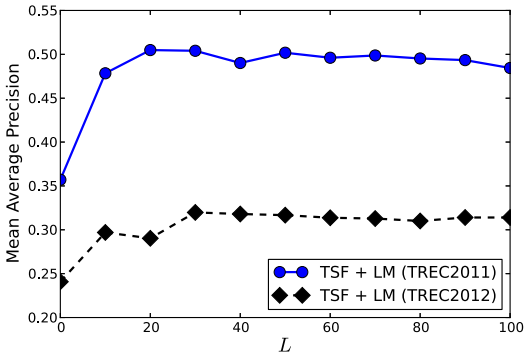

Figure 6: Sensitivity to the number of top retrieved tweets L used for tweet selection feedback. The x-axis shows the value of L. The y-axis shows the value of mean average precision over the TREC 2011 and 2012 microblog track topics, respectively.

Effect of Tweet Selection Feedback. We underscore the effectiveness of tweet selection feedback (TSF) comparing to initial search results (LM) in Figure 4. The bar plot shows the difference in average precision between LM and TSF on a query-by-query basis. Compared to relevance feedback methods without tweet selection feedback shown in Figure 1, TSF not only significantly improved search results over the initial search (see Table 1); it also improved the search performance of each topic without decreasing search performance over almost all topics. For example, Table 3 shows that TSF + LM improved results for about 97 topics, and decreased results for about 8 topics, whereas the results of relevance feedback methods without the use of TSF (RM, EXRM, TBRM, and QDRM) improved about 81–87 topics and decreased about 18–20 topics.

In addition, Figure 5 shows the results for relevance feedback after tweet selection (TSF + QDRM). Table 3 shows that TSF + RM, TSF + EXRM, TSF + TBRM, similarly to TSF + QDRM also improve retrieval performance for almost all topics without decreasing search performance compared to RM, EXRM and TBRM, which suggests that tweet selection feedback combined with PRF is effective to improve retrieval performance steadily. Particularly, we found that TSF + QDRM effectively uses search results refined by tweet selection feedback compared to other relevance feedback methods.

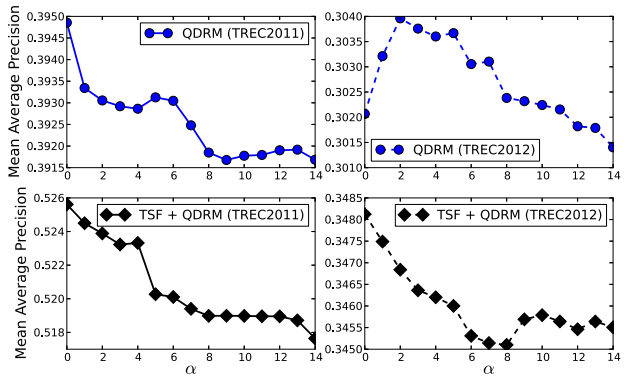

Figure 7: Sensitivity to the recency control parameter α used in QDRM over QDRM and TSF + QDRM at TREC 2011 (left-top and bottom) and QDRM and TSF + QDRM at TREC 2012 (right-top and bottom). The x-axis shows the values of α. The y-axis shows the value of mean average precision.

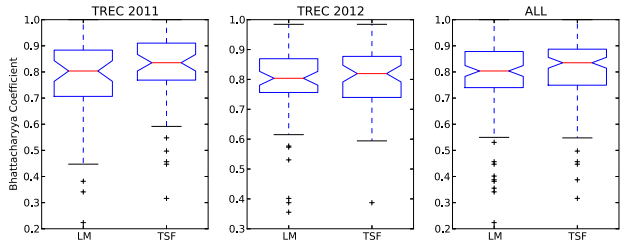

Figure 8: Bhattacharyya coefficient between temporal profiles of LM and TSF using the TREC 2011 and 2012 datasets.

Parameter Sensitivity. In our experiments, we selected a longest tweet among the top 30 tweets retrieved by LM (i.e. $L = 30$) and combined it with an original query as a new query for tweet selection feedback. We demonstrate in Figure 6 how the value of mean average precision (MAP) of TSF changes with different L parameters. Results showed that the performances of TSF + LM increase until $L = 30$, and become insensitive to L when L is large (e.g. $L \geq 30$) on both datasets. Those results suggest that the top ranked 30 tweets tend to contain a relevant tweet that can improve the retrieval performance via TSF, so that microblog users should read the top 30 tweets and select only a single relevant tweet among them when searching the Tweets2011 corpus effectively.

We also show the parameter sensitivity of α in QDRM. The parameter α controls the degree of the recency parameter over topics. Figure 7 shows the MAP values of QDRM and TSF + QDRM for $L = 30$, $M = 100$, $N = 10$, and $K = 20$ across different α values. It is readily apparent that the performance of QDRM and TSF + QDRM on both datasets is sharply decreasing when using large α values. Large α tempers the impact of temporal evidence because γ value tends to approach 0 (see Eqs. 14 and 17). The results suggest that query and document-dependent temporal evidence in QDRM is working. The optimal value of QDRM on TREC 2012 dataset is $\alpha = 2$, which indicates the effectiveness of recency. However, the difference of MAP values is slight. We assumed that this robustness results from the

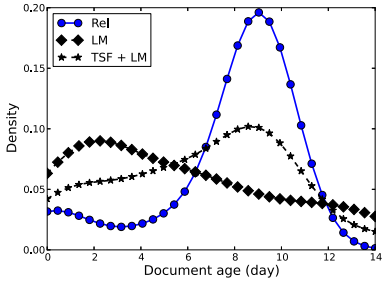

Figure 9: Temporal variations of a topic numbered MB042. The x-axis shows the document age from the query-time when query was issued to document time-stamp. The y-axis shows the kernel-estimated probability density for the document age. The blue line (Rel) shows estimates for relevant documents. Black lines (LM and TSF + LM) respectively show estimates of the top 30 retrieved documents by LM and TSF. High density indicates the period during which the topic was described actively.

short time span of the Tweets2011 corpus (about two weeks). It was also described earlier in the past work [9]. The optimal α values of TSF + QDRM on both datasets were 0, which means considering only query and document's temporal variation in temporal evidence ignoring recency is effect. We assumed that TSF was able to bring more accurate temporal distributions, so that the recency effect of QDRM was vanishingly small.

Temporal Analysis. We evaluate TSF from a temporal perspective. To demonstrate the improvement of estimation of temporal evidence, we compared the time-stamp distribution of relevant documents to that of the first 20 documents retrieved by a simple query likelihood model (LM) and tweet selection feedback (TSF + LM), respectively. Bhattacharyya coefficient is used as the similarity between two time-stamp distributions of retrieved documents. The higher value of the Bhattacharyya coefficient means that the IR system precisely estimates topic-related temporal evidence. Figure 8 shows the Bhattacharyya coefficients of LM and TSF + LM against relevant documents. To test the difference of the Bhattacharyya coefficient between LM and TSF + LM, we use two-sided Wilcoxon matched-pairs signed-ranks test with $p < 0.05$. Results show that the Bhattacharyya coefficient improved after TSF and the variance was smaller than LM. For example, the coefficient of TSF on TREC 2011 dataset significantly outperformed that of LM (from 0.7748 ± 0.1756 to 0.8071 ± 0.1566), but when using the TREC 2012 dataset, the difference is not statistically significant (from 0.7822 ± 0.1361 to 0.7988 ± 0.1112). The coefficient values of LM and TSF on both datasets (ALL) are 0.7789 ± 0.1553 and 0.8026 ± 0.1338, respectively. The difference is statistically significant. The point is that we can predict accurate temporal evidence using TSF.

Query Analysis. In Table 4, we display candidate words of query expansion for the topic *"Holland Iran envoy recall"* (MB042[6]) in RM, in QDRM, in TSF + RM, and in TSF +

[6]The news that "Dutch government is recalling its Tehran ambassador for consultations over the burial of executed Dutch–Iranian Sahra Bahram" was reported by BBC News on 7 February 2011.

Table 4: Expanded words for a topic numbered MB042: *"Holland Iran envoy recall"*.

RM		QDRM		TSF + RM		TSF + QDRM	
word	$P(w\|Q)$	word	$P(w\|Q)$	word	$P(w\|Q)$	word	$P(w\|Q)$
mubarak	0.057	mubarak	0.053	dutch	0.018	dutch	0.078
iranian	0.046	egypt	0.033	iranian	0.017	iranian	0.044
dutch	0.040	obama	0.031	bahrami	0.013	woman	0.032
says	0.036	rt	0.019	rt	0.012	bahrami	0.020
special	0.029	special	0.016	zahra	0.011	mubarak	0.019
row	0.028	stay	0.015	iranelection	0.009	zahra	0.017
stay	0.024	says	0.014	mubarak	0.009	drug	0.015
un	0.021	wisner	0.014	execution	0.008	hanging	0.014
news	0.020	jan25	0.011	woman	0.007	rt	0.013
egypt	0.018	frank	0.011	egypt	0.006	government	0.012

QDRM, showing improved results with TSF + QDRM. It is apparent that more topic-related words such as "dutch", "bahrami", and "iranian" appear in our approaches TSF + RM and TSF + QDRM. That is true because TSF selected *"Breaking: Dutch recall ambassador from #Iran over execution of Dutch–Iranian Zahra Bahrami, summon Iran Ambassador"* as a relevant document and used it for tweet selection feedback and refined lexical and temporal evidence for PRF. In addition, Figure 9 shows the kernel density estimate of the document age of topic MB042 using relevant tweets and top search results obtained using LM and TSF. From this figure, it is apparent that temporal variation approaches relevant estimates using TSF. Moreover, the word weights against a query, $P(w|Q)$, of topic-related words of TSF + QDRM are larger than that of TSF + RM. Results show that the average precision values of RM, QDRM, TSF + RM, and TSF + QDRM improved over the initial search LM (from 0.0490 to 0.0815, 0.0427, 0.8412, and 0.8478, respectively).

However, regarding *"Australian Open Djokovic vs. Murray"* (MB071), the average precision of relevance methods RM, QDRM improved over LM (from 0.5704 to 0.5809 and 0.6039, respectively) although that of TSF + RM, and TSF + QDRM decreased (from 0.5704 to 0.4450 and 0.4496, respectively). That is true because a tweet selected by TSF about this topic, *"Tomorrow is the Australian open tennis final for men, Andy Murray vs. Navok Djokovic Who's gonna win?? I'm a Murray fan so I say GO MURRAY!!"*, contains numerous topic-unrelated words. To improve the retrieval performance more using TSF, we must detect important concepts from this long query.

6. RELATED WORK

The proposed method combines the simple interactive query expansion method and the time-based PRF model for microblog search. In earlier work, term-based interactive query expansion methods were proposed, where users manually select topic-related words from suggested candidates as expansion words. They improved the retrieval performance [10, 35]. However, it is difficult to understand the context of suggested words. Their methods require cumbersome relevance judgments to select expansion words. In our method TSF, users must select no topic-related word. Instead, users merely read and select a single interesting tweet among initial search results.

Our approach is based on the notion of cluster-based information retrieval [13, 15, 16, 20, 36, 37] which uses clustering information to rank documents. Kurland and Lee [15] reranked documents using cluster information consisting of k nearest lexical similar documents. Liu and Croft [20] clustered all documents into several sets of similar documents using the k-means algorithm and used clusters for smoothing the language model with a global collection language model. Wei and Croft [37] proposed the document model using Latent Dirichlet Allocation to obtain clustering information for smoothing. Instead of smoothing for language models, Kalmanovich and Kurland [13] used cluster information with retrieved documents for creating an expanded query. In addition, Efron et al. [9] proposed a document expansion method based on the idea of Tao et al. [36], which smooths document language models by similar documents gathered with k nearest-neighbor. They submit documents as pseudo-queries to obtain similar documents, assuming that short documents such as a tweet tend to mention a single topic. Our approach differs from those in previous works in that we first made topic-related clusters by manually selecting a single tweet among initial search results and then submitting them to obtain similar documents. Consequently, our language modeling framework can easily reflect user intent. Additionally, we used this cluster for query expansion in the form of PRF as lexical and temporal evidence.

Microblog services often have real-time features by which many tweets are posted by crowds of people when a notable event occurs. Many reports have described studies about time-aware information retrieval methods for incorporating such real-time features. Dakka et al. [5] also proposed a general ranking mechanism integrating temporal properties into a language model identifying the important periods. Peetz et al. [29] proposed query modeling leveraging a temporal burst. Keikha et al. [14] proposed a time-based relevance model for improving blog retrieval. However, Dakka, Peetz, and Keikha's works cannot combine temporal properties of two types (recency and temporal variation) by topic. Li and Croft [18] incorporated recency into the language model framework for IR [17, 30]. Peetz et al. [28] tested many temporal document priors based on cognitive motivation for retrieving recent documents. However, their methods were unable to consider query-dependent recency. The query-dependent recency model was recently discussed in many works. Amati et al. [1] incorporated temporal recency into the document prior using survival function for microblog search. Massoudi et al. [22] proposed a query expansion method selecting words that are temporally closer to the query-time. Efron and Golovchinsky [8] proposed IR methods incorporating temporal properties into language modeling and showed their effectiveness for recency queries. Efron [7] also proposed a query-specific re-

cency ranking approach. In addition, Miyanishi et al. [26] combines recency and temporal variation based query expansion methods in response to query-dependent temporal property. However, they did not incorporate document-dependent temporal variation into their query expansion model. Our method takes account of lexical evidence weighted by temporal evidence related to each document while simultaneously considering recency.

7. CONCLUSION

In this paper, we proposed two-stage relevance feedback approaches for microblog search using tweet selection feedback and query-document dependent temporal relevance feedback methods. Our two-stage relevance feedback considerably improved retrieval performance with minimum user interaction. First, the user selects only one relevant tweet among top ranked initial search results and combines it with an original user query for tweet selection feedback (TSF), where the combined query is used for re-retrieving documents. Second, to improve search results further, a query-dependent relevance model QDRM is applied to top ranked re-retrieved documents.

TSF is a simple and effective approach to overcome the vocabulary mismatching problem and to improve microblog retrieval performance. Microblog documents are very short and tend to mention a single topic. TSF succeeds in exploiting the microblog feature. The user can quickly read and can readily select a relevant document among top re-retrieved search results that contain good words. A set of document time-stamps indicates the topic-related time. Using improved top search results for relevance feedback, we were able to improve search results using our proposed QDRM, which combines lexical and query-document dependent temporal evidence. Our two-stage relevance feedback framework can plug in any PRF method after TSF. We evaluated our approach using the Tweets2011 corpus with TREC 2011 and 2012 microblog datasets. The experimentally obtained results indicate that TSF markedly improves retrieval performance without decreasing over almost all queries. In addition, the proposed PRF method, QDRM, further considerably improved microblog search performance compared to established PRF methods. Although TSF is extremely effective for microblog search, TSF sometimes fails to outperform the initial search because of the redundancy of the tweet content, which contains meaningless words that sometimes degrade search results. In future work, we plan to refine tweet selection feedback method combined with an automatic key-concept extraction method for long queries [2].

8. REFERENCES

[1] G. Amati, G. Amodeo, and C. Gaibisso. Survival analysis for freshness in microblogging search. In *CIKM*, pages 2483–2486, 2012.

[2] M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In *SIGIR*, pages 491–498, 2008.

[3] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *SIGIR*, pages 243–250, 2008.

[4] J. Choi and W. B. Croft. Temporal models for microblogs. In *CIKM*, pages 2491–2494, 2012.

[5] W. Dakka, L. Gravano, and P. G. Ipeirotis. Answering general time-sensitive queries. *TKDE*, 24(2):220–235, 2012.

[6] F. Diaz. Integration of news content into web results. In *WSDM*, pages 182–191, 2009.

[7] M. Efron. Query-specific recency ranking: Survival analysis for improved microblog retrieval. In *#TAIA*, 2012.

[8] M. Efron and G. Golovchinsky. Estimation methods for ranking recent information. In *SIGIR*, pages 495–504, 2011.

[9] M. Efron, P. Organisciak, and K. Fenlon. Improving retrieval of short texts through document expansion. In *SIGIR*, pages 911–920, 2012.

[10] D. Harman. Towards interactive query expansion. In *SIGIR*, pages 321–331, 1988.

[11] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *TOIS*, 20(4):422–446, 2002.

[12] R. Jones and F. Diaz. Temporal profiles of queries. *TOIS*, 25(3), 2007.

[13] I. G. Kalmanovich and O. Kurland. Cluster-based query expansion. In *SIGIR*, pages 646–647, 2009.

[14] M. Keikha, S. Gerani, and F. Crestani. Time-based relevance models. In *SIGIR*, pages 1087–1088, 2011.

[15] O. Kurland and L. Lee. Corpus structure, language models, and ad hoc information retrieval. In *SIGIR*, pages 194–201, 2004.

[16] O. Kurland, L. Lee, and C. Domshlak. Better than the real thing? iterative pseudo-query processing using cluster-based language models. In *SIGIR*, pages 19–26, 2005.

[17] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR*, pages 120–127, 2001.

[18] X. Li and W. Croft. Time-based language models. In *CIKM*, pages 469–475, 2003.

[19] F. Liang, R. Qiang, and J. Yang. Exploiting real-time information retrieval in the microblogosphere. In *JCDL*, pages 267–276, 2012.

[20] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *SIGIR*, pages 186–193, 2004.

[21] Y. Lv and C. Zhai. Positional relevance model for pseudo-relevance feedback. In *SIGIR*, pages 579–586, 2010.

[22] K. Massoudi, M. Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. In *ECIR*, pages 362–367, 2011.

[23] D. Metzler, C. Cai, and E. Hovy. Structured event retrieval over microblog archives. In *HLT/NAACL*, pages 646–655, 2012.

[24] D. Metzler and W. B. Croft. Latent concept expansion using markov random fields. In *SIGIR*, pages 311–318, 2007.

[25] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *SIGIR*, pages 206–214, 1998.

[26] T. Miyanishi, K. Seki, and K. Uehara. Combining recency and topic-dependent temporal variation for microblog search. In *ECIR*, pages 331–343, 2013.

[27] I. Ounis, C. Macdonald, J. Lin, and I. Soboroff. Overview of the TREC-2011 microblog track. In *TREC*, 2011.

[28] M.-H. Peetz and M. de Rijke. Cognitive temporal document priors. In *ECIR*, pages 318–330, 2013.

[29] M. H. Peetz, E. Meij, M. de Rijke, and W. Weerkamp. Adaptive temporal query modeling. In *ECIR*, pages 455–458, 2012.

[30] J. Ponte and W. Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281, 1998.

[31] J. J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System*, pages 313–323, 1971.

[32] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM*, pages 623–632, 2007.

[33] I. Soboroff, I. Ounis, and J. Lin. Overview of the TREC-2012 microblog track. In *TREC*, 2012.

[34] T. Strohman, D. Metzler, H. Turtle, and W. Croft. Indri: a language model-based search engine for complex queries. In *ICIA*, 2005.

[35] B. Tan, A. Velivelli, H. Fang, and C. Zhai. Term feedback for information retrieval with language models. In *SIGIR*, pages 263–270, 2007.

[36] T. Tao, X. Wang, Q. Mei, and C. Zhai. Language model information retrieval with document expansion. In *HLT/NAACL*, pages 407–414, 2006.

[37] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *SIGIR*, pages 178–185, 2006.

[38] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM*, pages 403–410, 2001.

[39] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *TOIS*, 22(2):179–214, 2004.