Time-Aware Latent Concept Expansion for Microblog Search

Taiki Miyanishi and Kazuhiro Seki and Kuniaki Uehara
Graduate School of System Informatics, Kobe University
1-1 Rokkodai, Nada, Kobe 657-8501, Japan

Abstract

Incorporating the temporal property of words into query expansion methods based on relevance feedback has been shown to have a significant positive effect on microblog search. In contrast to such word-based query expansion methods, we propose a concept-based query expansion method based on a temporal relevance model that uses the temporal variation of concepts (e.g., terms and phrases) on microblogs. Our model naturally extends an extremely effective existing concept-based relevance model by tracking the concept frequency over time. Moreover, the proposed model produces important concepts that are frequently used within a particular time period associated with a given topic, which better discriminate between relevant and non-relevant microblog documents than words. Our experiments using a corpus of microblog data (Tweets2011 corpus) show that the proposed concept-based query expansion method improves search performance significantly, especially for highly relevant documents.

1 Introduction

Time plays an important role in retrieving relevant and informative microblogs because of the real-time feature of microblog documents (Efron and Golovchinsky 2011; Efron, Organisciak, and Fenlon 2012; Lin and Efron 2013; Peetz et al. 2012). Particularly, query expansion methods based on relevance feedback incorporating the temporal property of words into their models have been demonstrated as effective for improving microblog search performance (Choi and Croft 2012; Massoudi et al. 2011; Metzler, Cai, and Hovy 2012; Miyanishi, Seki, and Uehara 2013a; 2013b). These time-based query expansion methods mainly use word frequency in pseudo-relevant documents as lexical information and temporal variations of word frequency as temporal information.

However, such word-based pseudo-relevance feedback (PRF) methods result in limited retrieval effectiveness for retrieving highly relevant documents. The fundamental reason is that words have semantic ambiguity. Furthermore, word frequency often fails to indicate the exact time-ranges in which crowds of people are interested (Miyanishi, Seki, and Uehara 2013a).

To overcome the shortcomings of word-based IR, several researchers have recently proposed unsupervised or supervised concept importance weighting methods (Bendersky and Croft 2008; 2012; Bendersky, Metzler, and Croft 2010; 2011; 2012; Lang et al. 2010; Lease 2009; Metzler and Croft 2005; 2007) because concepts (e.g., terms and phrases) generally have more discriminative power than words. However, the existing concept-based IR models do not consider time, which is an important factor for microblog search, because these methods are mainly used for Web searches, which require almost no temporal information. Therefore, the open question we are tackling is the weighting of concepts effectively using temporal information.

To address this question, we propose a novel concept weighting scheme based on the temporal relevance model for query expansion. The proposed model extends a state-of-the-art concept weighting approach, called Latent Concept Expansion (LCE) (Metzler and Croft 2007), from a temporal perspective. We call this method time-aware latent concept expansion, which provides a unified framework for weighting concepts using both lexical and temporal information.

To clarify differences between the existing methods and the proposed one, Table 1 contrasts words and concepts suggested by a standard word-based PRF method (Lavrenko and Croft 2001), wRM, a standard concept-based lexical PRF method, cTRM (Lexical) that is equal to LCE (Metzler and Croft 2007), and our proposed concept-based temporal PRF method using only temporal information, cTRM (Temporal), for a topic numbered MB044: “White House spokesman replaced” used in the TREC microblog track. This topic is related to the news that Jay Carney, who had

<table>
<thead>
<tr>
<th>wRM</th>
<th>cTRM (Lexical)</th>
<th>cTRM (Temporal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>jay</td>
<td>jay</td>
<td>carney</td>
</tr>
<tr>
<td>carney</td>
<td>carney</td>
<td>jay</td>
</tr>
<tr>
<td>qantas</td>
<td>qantas</td>
<td>press secretary</td>
</tr>
<tr>
<td>new</td>
<td>new spokesmen</td>
<td>jay carney</td>
</tr>
<tr>
<td>obama</td>
<td>new</td>
<td>biden spokesmen</td>
</tr>
</tbody>
</table>

Table 1: Example of expanded words and concepts for a topic “White House spokesman replaced” from a word-based PRF (wRM) and a concept-based temporal one (cTRM).
been the chief spokesman for Vice President Joseph R. Biden Jr., took over as White House Press Secretary. Table 1 clarifies that the word-based PRF method wRm suggests topic-related words jay and carney. However, jay and carney often retrieve irrelevant documents because these words appear in many documents. In contrast, concept-based methods cTRM (Lexical) and cTRM (Temporal) suggest exact topic-related concepts: new spokesman, press secretary, and jay carney. It is particularly interesting that in this case that the PRF method using only temporal information, cTRM (Temporal), suggests more topic-related and different concepts than cTRM (Lexical). Therefore, we assume that our temporal PRF method, cTRM, integrating lexical and temporal information for selecting topic-related concepts will be more effective than a PRF method using only lexical information (e.g., LCE) as well as the standard word-based PRF method.

This paper has two primary contributions. First, we describe a novel time-based relevance model. Our model provides a flexible framework for selecting important words and concepts associated with a specified time period. This framework is a natural extension of standard word and concept weighting schemes (Lavrenko and Croft 2001; Metzler and Croft 2007) from a temporal perspective. Second, we carry out a detailed empirical evaluation which demonstrates the state-of-the-art effectiveness of the proposed model on a standard test collection for microblog search (Tweets2011 corpus). Our evaluation shows that the proposed PRF using multi-term concepts is particularly beneficial for retrieving highly relevant documents.

The remainder of the paper is organized as follows: in Sec. 2 we survey related work. Sec. 3 describes details of the proposed concept-based temporal relevance model. Experimental settings and results are presented in Sec. 4. Finally, Sec. 5 presents a summary of this work and conclusions.

2 Related Work

The proposed time-aware latent concept expansion is an algorithm for expanding an original query with multi-term concepts that are frequently used within a topically relevant time period. It derived from the notion of time-aware information retrieval and concept-based information retrieval. We describe these related work below.

2.1 Time-Aware Information Retrieval

People search microblog documents to find temporally relevant information, such as breaking news and real-time content (Teevan, Ramage, and Morris 2011), so that temporal properties (e.g., recency and temporal variations) are important factors for retrieving such information. For detecting temporally relevant information, many studies have incorporated temporal properties into their respective frameworks. Li and Croft (2003) incorporated recency into the language model framework for information retrieval (IR) (Lavrenko and Croft 2001; Ponte and Croft 1998). Efron and Golovchinsky (2011) also incorporated temporal properties, especially recency, into language model smoothing. Dakka et al. (2012) proposed a general rank-
3 Proposed method

The proposed query expansion method based on a PRF model builds on language modeling frameworks (a query likelihood model) for IR. Thus, we first introduce the query likelihood model and the relevance model based on language modeling frameworks. Then, we describe the proposed concept-based temporal relevance model for query expansion.

3.1 Language Model for Information Retrieval

The query likelihood model (Ponté and Croft 1998) incorporates the assumption that the probability of a query \( Q \) is generated by the word probabilities on a document \( D \). All documents are ranked in order of their probability of relevance or usefulness, which is defined as \( P(D|Q) \). The posterior probability of a document \( P(D|Q) \) by Bayes’ rule becomes

\[
P(D|Q) \propto P(D)P(Q|D),
\]

where \( P(Q|D) \) denotes the query likelihood on the given document and \( P(D) \) stands for 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 \( P(Q|D) \), where the query \( Q \) consists of \( n \) query terms \( q_1, q_2, \ldots, q_n \), as

\[
P(Q|D) = \prod_{i=1}^{n} P(q_i|D),
\]

where \( P(q_i|D) \) is the probability of an \( i \)-th query term \( q_i \) under the word distribution for document \( D \). The maximum likelihood estimator of \( P(q_i|D) \) is \( P_{ml}(w|D) = \frac{f(w;D)}{\sum_{w \in \mathcal{V}} f(w;D)} \). Therein, \( f(w;D) \) denotes the number of word counts of \( w \) in document \( D \), \( \sum_{w \in \mathcal{V}} f(w;D) \) is the number of words in \( D \) where \( \mathcal{V} \) is the set of all words in the vocabulary. In most cases, this probability is applied to smoothing to temper over-fitting using a given collection. Among numerous smoothing methods, the following Dirichlet smoothing (Zhai and Lafferty 2004) is often used.

\[
P(w|D) = \frac{|D|}{|D| + \mu} P_{ml}(w|D) + \frac{\mu}{|D| + \mu} P(w|C),
\]

where \( \mu \) is the Dirichlet prior and \( P(w|C) \) is a uni-gram language model in a corpus \( C \). Smoothing the maximum likelihood estimator of the uni-gram language model improves the estimated probabilities.

3.2 Word-based Relevance Model

In this section, we introduced existing PRF methods using only lexical information of words and concepts. Lavrenko and Croft (2001) incorporated relevance feedback into language modeling frameworks. They estimated a relevance model, \( P(w|R) \), using a joint probability of observing the expanded word \( w \) together with query terms in query \( Q \), assuming that the word \( w \) was sampled in the same way as the query terms from a distribution \( R \). That relevance model weights words \( w \) according to the following.

\[
P(w|R) \approx P(w|Q) = \sum_{D \in R} P(w,D|Q)
\]

\[
= \frac{1}{Z} \sum_{D \in R} P(D)P(w,Q|D)
\]

\[
\propto \sum_{D \in R} P(D)P(w|D) \prod_{i} P(q_i|D), \quad (2)
\]

where \( R \) is a set of relevant or pseudo-relevant document for query \( Q \) and where \( Z = \sum_{w \in V} \sum_{D \in R} P(w,D,q) \) is a normalization factor. When using the top \( M \) retrieved documents by the query \( Q \) for \( R \), this approach is called pseudo-relevance feedback. In addition, for query expansion, words \( w \) are ordered in descending order of \( P(w|Q) \) in Eq. 2. Then, the top \( k \) words are added to the original user query. Recall that this relevance model uses only word frequency.

3.3 Concept-based Relevance Model

To model query concepts through term dependencies for PRF, Metzler and Croft (2007) proposed the concept-based PRF method called LCE, which generates single and multi-term concepts that are related topically to an original query. These concepts are defined as latent concepts. To represent term-dependencies in a query and documents, LCE mainly uses the notion of Markov random field (Metzler and Croft 2005). Using LCE, users can automatically formulate the concepts a user has in mind, but which the user did not explicitly express in the query. The goal of LCE is to recover these latent concepts given some original query. As described in this paper, we used the simplified LCE proposed by Bendersky et al. (2011) to assess the effectiveness of several components between baselines and our proposed approach. Their LCE weights a latent concept extracted from pseudo-relevant documents \( R \) (top \( M \) retrieved documents) as follows:

\[
S_{LCE}(c,Q) \propto \sum_{D \in R} \exp[\gamma_1 \phi_1(Q,D) + \gamma_2 \phi_2(c,D) + \gamma_3 \phi_3(c,C)], \quad (3)
\]

where \( \phi_1(Q,D) \) is a matching function between a document \( D \) and concepts in a query \( Q \), \( \phi_2(c,D) \) is the matching function between a concept \( c \) and the document \( D \), and \( \phi_3(c,C) \) is the the matching function of the concept \( c \) in the corpus \( C \).

Moreover, we assume that the given query consisting of query concepts \( c_1, c_2, \ldots, c_m \) in \( Q \) and the candidates of an expanded concept \( c \) in pseudo-relevant documents are sampled identically and independently from a concept uni-gram
distribution of \( R \), namely, assuming the bag-of-concepts. When \( \phi_1(Q, D) = \log P(Q|D) \), \( \phi_2(c, D) = \log P(c|D) \), \( \phi_3(c, C) = 0 \), and \( \gamma_1 = \gamma_2 = \gamma_3 = 1 \), we obtain the score function of a concept \( c \) in response to query \( Q \) as

\[
S_{c, RM}(c, Q) \propto \sum_{D \in R} P(D) P(c|D) \prod_{i} P(\hat{q}_i|D), \quad (4)
\]

where \( \hat{q}_i \) is \( i \)-th query concept in query \( Q \). This PRF model drops the penalty of the inverse collection frequency of the concept in the corpus from Eq. 3.\(^1\) In addition, the expansion of Eq. 4 is similar to the word-based PRF model in Eq. 2. Unlike the word-based PRF that uses only words, concept-based PRF in Eq. 4 can use multi-term concepts as well as single words. However, existing word-based and concept-based methods cannot use temporal information such as document time-stamps, which are important features for microblog search.

3.4 Concept-based Temporal Relevance Model

Microblog services often have real-time features for which many microblogs are posted by crowds of people when a notable event occurs (Sakaki, Okazaki, and Matsuo 2010). Many reports have described the effectiveness of incorporating such real-time features into PRF methods for microblog search (Choi and Croft 2012; Massoudi et al. 2011; Miyanishi, Seki, and Uehara 2013a; 2013b). Therefore, we propose a concept-based PRF method that combines lexical and temporal information of concepts.

We assume that the proposed concept-based relevant model \( P(c|R) \) derives from both lexical and temporal information sources. Therefore, we have

\[
P(c|Q) = \sum_{D_1 \in R_1} \sum_{D_2 \in R_2} P(c, D_1, D_2|Q) = \sum_{D_1 \in R_1} \sum_{D_2 \in R_2} P(D_1|c, D_2, Q) P(c, D_2|Q), \quad (5)
\]

where \( D_1 \) denotes a document from pseudo-relevant documents \( R_1 \) and \( D_2 \) denotes each time (a day in our case) in \( R_2 \). Then, as with the work by Efron and Golovchinsky (2011), we apply the simple assumption that the temporal information \( D_2 \) is independent of the lexical information \( D_1 \), so that \( D_2 \) is dropped from the conditional probability in Eq. 5. Therefore, we have

\[
P(c|Q) = \sum_{D_1 \in R_1} P(D_1|c, Q) \sum_{D_2 \in R_2} P(c, D_2|Q) = \frac{1}{P(c|Q)} \sum_{D_1 \in R_1} P(c, D_2|Q) \sum_{D_2 \in R_2} P(c, D_2|Q) \propto \frac{1}{P(c|Q)} \sum_{D_1 \in R_1} P(D_1) P(c, Q|D_1) \sum_{D_2 \in R_2} P(D_2) P(c, Q|D_2)
\]

Then, following the notion of bag-of-features, we assume that query concepts \( \hat{q}_1, \hat{q}_2, \ldots, \hat{q}_m \) and concept \( c \) for query expansion are sampled identically and independently from a lexical distribution of pseudo-relevant documents, \( R_1 \), and a time distribution of ones, \( R_2 \) (top \( N \) retrieved documents). We have

\[
P(c|Q) \propto \frac{1}{P(c|Q)} \sum_{D_1 \in R_1} \sum_{D_2 \in R_2} P(D_1) P(c, D_2|Q) \prod_{i} P(\hat{q}_i|D_1) \cdot \prod_{j} P(\hat{q}_j|D_2)
\]

where \( P(c|D_1) \) and \( P(\hat{q}_i|D_2) \) denote the probability of concept occurrence in document \( D_1 \); \( P(c|D_2) \) and \( P(\hat{q}_j|D_2) \) denote the probability of concept occurrence at time \( t \). Then, because \( P(c|Q) \) is a non-negative function, we have the score function that ranks a concept \( c \) in response to query \( Q \) as

\[
S_{c, TRM}(c, Q) \overset{\text{rank}}{=} \left\{ \sum_{D_1 \in R_1} \sum_{D_2 \in R_2} P(D_1) P(c, D_2|Q) \prod_{i} P(\hat{q}_i|D_1) \cdot \prod_{j} P(\hat{q}_j|D_2) \right\}^{1/2}, \quad (6)
\]

Here \( P(D_1) \) and \( P(D_2) \) are uniform over all the distributions in \( D_1 \) and \( D_2 \). The value of \( P(c|D_1) \prod_j P(\hat{q}_j|D_2) \) increases when the candidate concept \( c \) and query concepts \( \hat{q}_1, \hat{q}_2, \ldots, \hat{q}_m \) were described together simultaneously in a range. Using the probabilities of concept occurrence \( P(c|D_1) \) derived from document time-stamps of pseudo-relevant documents \( R_1 \), this PRF model represents real-time feature of a given topic in microblogging services. In addition, because \( P(c|D_1) \prod_j P(\hat{q}_j|D_2) \) is equal to a factor of the standard concept-based PRF method, LCE (see Eq. 4), Eq. 6 is obtained for the product of lexical concept information and a temporal one. Figure 1 clarifies the difference between the existing concept-based relevance modeling (LCE) and the proposed concept-based temporal relevance modeling.

To improve our estimates for \( P(c|D_1) \), we also use Dirichlet smoothing as with the standard query likelihood model in Eq. 1 because the value of query likelihood \( \prod_j P(\hat{q}_j|D_2) \) becomes 0 when a query concept \( \hat{q}_i \) does not appear over time in \( R_2 \). We have

\[
P(c|D_1) = \frac{|D_1|}{|D_1| + \mu_t} P_{ml}(c|D_1) + \frac{\mu_t}{|D_1| + \mu_t} P(c|C), \quad (7)
\]

Figure 1: Graphical model representations of concept-based relevance modelling (left) and the proposed concept-based temporal relevance modelling (right).

---

\(^1\)Because the concept frequencies contribute little to the significant improvements in retrieval performance (Macdonald and Ounis 2010), we set \( \phi_3(c, C) = 0 \).
that we used in our experiments.

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>#Topics</th>
<th>Topic Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC 2011</td>
<td>allrel</td>
<td>49</td>
<td>1-49</td>
</tr>
<tr>
<td></td>
<td>highrel</td>
<td>33</td>
<td>1, 10-30, 32, 36-38, 40-42, 44-46, 49</td>
</tr>
<tr>
<td>TREC 2012</td>
<td>allrel</td>
<td>59</td>
<td>51-75, 77-110</td>
</tr>
<tr>
<td></td>
<td>highrel</td>
<td>56</td>
<td>51, 52, 54-68, 70-77, 77-104, 106-110</td>
</tr>
</tbody>
</table>

where \( \hat{P}_{ml}(c|D_t) = \frac{f(c,D_t)}{\sum_{c \in \mathcal{V}_c} f(c,D_t)} \), \( \mathcal{V}_c \) is the set of all concepts in the vocabulary of concepts, \( f(c,D_t) \) is the frequency of concept \( c \) at time \( t \), \( |D_t| \) is the total number of concepts at time \( t \), \( \mu_t \) is a parameter for smoothing, and \( P(c|C) \) is the probability of concept \( c \) occurrence in the corpus \( C \). Finally, we rank candidate concepts in descending order of the association score \( S_{cTRM}(c,Q) \) and use the top \( k \) concepts for query expansion.

### 4 Evaluation

This section describes the details of our experimental evaluation. First, in Sec. 4.1, we describe the experimental setup used for the evaluation. Then, in Sec. 4.2, we show baselines to compare our proposed method. Sec. 4.3 explains evaluation metrics and a statistical test for our evaluation. In Sec. 4.4, we compares the performance of the temporal query expansion to the performance of several standard atemporal retrieval methods. Finally, Sec. 4.5 provides additional experiments to discuss various aspects of the proposed method.

#### 4.1 Experimental Setup

**Evaluation data** We evaluated our proposed method using the test collection for the TREC 2011 and 2012 microblog track (Tweets2011 corpus\(^3\)). This collection consists of about 16 million tweets sampled between January 23 and February 8, 2011, for 110 search topics. Fig. 2 presents an example topic from the TREC 2011 and 2012 microblog tracks. In the figure, \( \langle \text{num} \rangle \) is a topic number, \( \langle \text{title} \rangle \) is a user query, and \( \langle \text{querytime} \rangle \) is the query-time when the query was issued. In our experiments, we use \( \langle \text{title} \rangle \) as a test query which is the official query used in the TREC 2011 and 2012 microblog track.

To evaluate any IR system, relevance judgment is applied to the whole tweet set of each topic. The relevance levels are categorized into irrelevant (labeled 0), minimally relevant (labeled 1), and highly relevant (labeled 2). We separately evaluated our method with respect to allrel and highrel query sets: allrel has both minimally relevant and highly relevant tweets as relevant documents and highrel has only highly relevant tweets. Table 2 summarizes topic numbers that we used in our experiments.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>#Topics</th>
<th>Topic Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC 2011</td>
<td>allrel</td>
<td>49</td>
<td>1-49</td>
</tr>
<tr>
<td></td>
<td>highrel</td>
<td>33</td>
<td>1, 10-30, 32, 36-38, 40-42, 44-46, 49</td>
</tr>
<tr>
<td>TREC 2012</td>
<td>allrel</td>
<td>59</td>
<td>51-75, 77-110</td>
</tr>
<tr>
<td></td>
<td>highrel</td>
<td>56</td>
<td>51, 52, 54-68, 70-77, 77-104, 106-110</td>
</tr>
</tbody>
</table>

\(^3\)http://www.lemurproject.org/indri/

**Microblog search settings** We indexed tweets posted before the specific time associated with each topic by the Indri search engine\(^3\) with the following setting. All queries and tweets were stemmed using the Krovetz stemmer (Krovetz 1993) without stop-word removal. They were case-insensitive. We built an index for each query. This index was created to simulate a realistic real-time search setting, where no future information is available when a query is issued.

To retrieve documents, we used a basic query likelihood model with Dirichlet smoothing (Zhai and Laﬀerty 2004) (we set smoothing parameter \( \mu = 2500 \) similar to Efron’s work (2012)) implemented by the Indri search engine (Strohman et al. 2005) as the language model for IR (LM) and all PRF methods used this LM as initial search results. For temporal smoothing parameter \( \mu_t \) in Eq. 7, we set \( \mu_t = 150 \) when retrieving documents for allrel queries, and let \( \mu_t = 350 \) for highrel based on results of a pilot experiment. In addition, instead of direct estimation of \( P(c|C) \), we used \( P(c|C) \approx \frac{df(c)}{N} \), where \( df(c) \) is the document frequency of concept \( c \) and \( N \) is the total number of documents in the corpus because it can be expensive to calculate the number of documents containing a pair of query terms. Even though \( df(c)/N \) is different from \( P(c|C) \), we coordinate the difference with the smoothing parameter \( \mu_t \). The sensitivity of a parameter \( \mu_t \) is discussed in Sec. 4.5.

We filtered out all non-English retrieved tweets using a language detector with infinity-gram, called ldig\(^4\). Retweets\(^5\) were regarded as irrelevant for evaluation in the TREC Microblog track (Ounis et al. 2011; Soboroff, Ounis, and Lin 2012); however, we used retweets except in a final ranking of tweets because a set of retweets is a good source that might contain topic-related words for improving Twitter search performance (Choi and Croft 2012). In accordance with the track’s guidelines, all tweets with http status codes of 301, 302, 403, and 404 and all retweets including the string “RT” at the beginning of the tweet were removed from the final ranking. Finally, we used the top 1000 results for evaluation.

\(^4\)https://github.com/shuyo/ldig

\(^5\)Tweets re-posted by another user to share information with other users

Table 3: Summary of evaluated retrieval methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Lexical</th>
<th>Temporal</th>
<th>Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>wRM</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cRM</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>wTRM</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>cTRM</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Figure 2: Example topic from the TREC microblog track.
4.2 IR Models

Baselines First, we introduce the setting of the proposed PRF method. Then we describe baselines to validate the effectiveness of each component in our proposed method.

The concept-based method uses the combination of one or two words as a candidate concept. All concepts are extracted from tweets based on sequential dependence, which assumes that dependence exists between adjacent query terms (Metzler and Croft 2005). Previous PRF methods also use this sequential dependence model (Bendersky, Metzler, and Croft 2010; Metzler and Croft 2007) because this model has consistently demonstrated state-of-the-art retrieval effectiveness in Web search. Although we use the sequential dependence model in this study, our model uses no independence structure. In addition, we used two types of concept such as #1(·) and #uw8(·), where #1(·) denotes an ordered window in which words must appear adjoinly ordered and #uw8(·) denotes an unordered window in which all words must appear within a window of 8 terms in any order. We denote the proposed PRF method combining lexical and temporal information of concepts as cTRM.

Moreover, to assess the effectiveness of incorporating concept into the retrieval model, we also proposed a word-based temporal relevance model, wTRM, that incorporates lexical and temporal information of words into its relevance model. wTRM uses only a single word as a concept in Eq. 6: wTRM does not consider multi-term concepts that combine more than two words. We compare this model wTRM to cTRM that uses lexical and temporal information of any concept.

To assess our proposed method cTRM, we prepared two baseline methods. The first baseline, wRM, uses a standard relevance feedback using only lexical information of words (Lavrenko and Croft 2001). In other words, wRM uses only word information. It does not consider multiple term concepts and temporal information. Note that cTRM reduces to wRM when the number of pseudo-relevant documents from temporal perspective, R_t, is 0 and all using concepts are single words (see Eqs. 2 and 6).

Our second baseline, cRM, uses pseudo-relevance feedback with lexical information of concepts. This method is equivalent to Latent Concept Expansion (LCE) (Metzler and Croft 2007), except for some points. To validate the effectiveness of concept’s temporal information, we use simplified LCE in Eq. 4. This PRF model drops the penalty of the inverse collection frequency of the concept in corpus from Bendersky’s LCE in Eq. 3. Both cRM and cTRM can use any concept. However, cRM differs from cTRM in that cRM does not consider temporal information such as R_t.

Table 3 summarizes the choice of concepts and pseudo-relevance information sources used by our methods and baselines. For instance, it is apparent from Table 3 that cRM and cTRM share the same concept types, but differ in the type of pseudo-relevant documents for concept reweighting. Note that the PRF methods using only lexical information, wRM and cRM, are strong baselines. The PRF methods using lexical and temporal information, wTRM and cTRM, are our proposed approaches.

```
#weight( λ1 #combine(bbc world service staff cuts) λ2 #weight( c1 #1(service outlines) c2 #uw8(bbc outlines) c3 outlines ... ck #1(weds bbcworldservice)))
```

Figure 3: Example of query expansion of topic “BBC World Service staff cuts” from TREC microblog track queries.

Query expansion For all PRF methods, we select candidate words or concepts among the top M tweets retrieved using the original query after removing the uniform resource locators (URLs), and user names starting with ‘@’ or special characters (!, @, #, ', "”, etc.). All query terms, candidates of words and concepts, and tweets are decapitalized. The candidates of words and concepts include no stop-words prepared in the Indri search engine. Then, we select k words or concepts among candidates in descending order of the word or concept weighting score, such as \( S_wRM(c, Q) \) or \( S_MTRM(c, Q) \). We use the normalized score for concept weighting. For example, the weight of i-th concept is \( c_i = \frac{S_MTRM(c_i, Q)}{\sum_j S_MTRM(c_j, Q)} \) when using cTRM. Finally, we combined the expanded concepts of PRF with their weight and the original query as an expanded query. They were weighted with 1:1. Fig. 3 shows an example of query expansion we used. In our study, we set \( \lambda_1, \lambda_2 = 0.5 \).

For wTRM and cTRM, we tuned parameters: the number of pseudo-relevant documents as temporal information (i.e., N). For all methods, we also tuned their parameters: the number of pseudo-relevance feedback documents (i.e., M) and the number of expansion words (i.e., k). Values of these parameters were optimized for best performance of Mean Average Precision (MAP) on training data because MAP is a stable measure. For example, we tuned parameters of the IR model using TREC 2012 microblog track dataset and tested it with TREC 2011 microblog dataset. In contrast, we trained the model using the TREC 2012 dataset and tested it on the TREC 2011 dataset. The sensitivity of some parameters such as N in wTRM and cTRM and the number of words or concepts used for query expansion, k, is discussed in Sec. 4.5.

4.3 Evaluation Measure

The goal of our system is to return a ranked list of tweets using relevance feedback methods. To evaluate retrieval effectiveness, we used average precision (AP), R-Precision (Rprec), and binary preference (bpref). AP is the mean of the precision scores obtained after each relevant document is retrieved. Rprec is that precision after R documents have been retrieved where R is the number of relevant document for the given topic. Bpref considers whether relevant documents are ranked above irrelevant ones. AP and Rprec have lower error rates than Precision (Buckley and Voorhees 2000). Bpref is more robust evaluation measure than AP when using incomplete relevance data (Buckley and Voorhees 2004).
Table 4: Performance comparison of the word-based PRF methods. Superscripts $\alpha$, $\beta$, and $\gamma$ respectively denote statistically significant improvements over LM, wRM, and wTRM. The best result per column is marked by boldface.

<table>
<thead>
<tr>
<th>Method</th>
<th>AP</th>
<th>Rprec</th>
<th>bpref</th>
<th>AP</th>
<th>Rprec</th>
<th>bpref</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>0.2936</td>
<td>0.3313</td>
<td>0.3103</td>
<td>0.2130</td>
<td>0.2286</td>
<td>0.1933</td>
</tr>
<tr>
<td>wRM</td>
<td>0.3502$^{\alpha}$</td>
<td>0.3868$^{\alpha}$</td>
<td>0.3594$^{\alpha}$</td>
<td>0.2473$^{\alpha}$</td>
<td>0.2537</td>
<td>0.2242</td>
</tr>
<tr>
<td>wTRM</td>
<td>0.3726$^{\beta}$</td>
<td>0.4089$^{\beta}$</td>
<td>0.3872$^{\beta}$</td>
<td>0.2580$^{\beta}$</td>
<td>0.2705$^{\beta}$</td>
<td>0.2361$^{\beta}$</td>
</tr>
</tbody>
</table>

Table 5: Performance comparison of the concept-based PRF methods. Superscripts $\alpha$, $\beta$, and $\gamma$ respectively denote statistically significant improvements over LM, cRM, and cTRM. Best result per column is marked by boldface.

<table>
<thead>
<tr>
<th>Method</th>
<th>AP</th>
<th>Rprec</th>
<th>bpref</th>
<th>AP</th>
<th>Rprec</th>
<th>bpref</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>0.2936</td>
<td>0.3313</td>
<td>0.3103</td>
<td>0.2130</td>
<td>0.2286</td>
<td>0.1933</td>
</tr>
<tr>
<td>cRM</td>
<td>0.3385$^{\alpha}$</td>
<td>0.3725$^{\alpha}$</td>
<td>0.3479$^{\alpha}$</td>
<td>0.2511$^{\alpha}$</td>
<td>0.2696$^{\alpha}$</td>
<td>0.2356$^{\alpha}$</td>
</tr>
<tr>
<td>cTRM</td>
<td>0.3644$^{\beta}$</td>
<td>0.4058$^{\beta}$</td>
<td>0.3825$^{\beta}$</td>
<td>0.2694$^{\beta}$</td>
<td>0.2770$^{\beta}$</td>
<td>0.2527$^{\beta}$</td>
</tr>
</tbody>
</table>

To validate the retrieval effectiveness, we discuss the statistical significance of results obtained using a two-sided Fisher’s randomization test (Smucker, Allan, and Carterette 2007), which is a non-parametric statistical significance test that does not assume the specific distribution. We used a Perl implementation for the randomization test\(^9\) with 100,000 permutations and $p < 0.05$ through this paper.

4.4 Experimental Results

To assess the effectiveness of our proposed methods wTRM and cTRM, we compared wTRM and cTRM using standard PRF methods: wRM and cRM.

Comparison of word-based PRF methods Table 4 compares the retrieval effectiveness of the initial search (LM) and the word-based PRF method using only lexical information (Lavrenko and Croft 2001) (wRM) to the retrieval effectiveness of word-based PRF method using lexical and temporal information (wTRM), both for allrel and highrel queries. It is apparent from Table 4 that both wRM and wTRM markedly outperform the initial search LM on both measures across both query sets. In particular, wTRM improved search results with statistical significance in all cases. Moreover, wTRM outperformed the standard word-based relevance model wRM in terms of all evaluation measures across both query sets. The difference in AP and bpref for allrel queries was statistically significant, which suggests that incorporating temporal information through our model using single words as concepts is important for retrieving topical relevant microblogs.

Comparison of concept-based PRF methods Table 5 compares the retrieval effectiveness of LM and the concept-based PRF method using only lexical information (Bendersky, Metzler, and Croft 2011) (cRM) to the retrieval effectiveness of concept-based PRF method using lexical and temporal information (cTRM), both for allrel and highrel queries. Table 5 clarifies that both cRM and cTRM markedly outperform the initial search LM on both measures across both query sets with statistical significance as with word-based approaches: wRM and wTRM. Moreover, cTRM outperformed the standard concept-based PRF method cRM in terms of all evaluation measures across both query sets. Particularly, the differences in Rprec and bpref for using allrel queries and in AP for using highrel queries was statistically significant. The results suggest two findings. First, latent concept expansion for pseudo-relevance feedback, which uses multi-term concepts for query expansion, is effective for microblog search. This result is consistent with previous work (Metzler and Cai 2011). Second, temporal information of concepts for PRF method is an important factor for retrieving topically relevant microblog documents, so that the proposed cTRM consistently outperformed the state-of-the-art latent concept expansion method, cRM.

Comparison to the standard lexical PRF method This section presents a comparison of cTRM with a standard word-based PRF method (wRM). Table 6 compares the retrieval effectiveness of the standard word-based lexical PRF method (wRM) to the retrieval effectiveness of concept-based temporal PRF method cTRM, both for allrel and highrel queries. Table 6 clarifies that cTRM outperformed wRM in terms of all evaluation measures across both allrel and highrel query sets. Particularly, the differences in AP and bpref for highrel queries were statistically significant, whereas there are no significant differences between wRM and wTRM for highrel. The results suggest the combination of using a concept instead of single word for query expansion and using a temporal information of concepts for pseudo-relevance feedback is effective to retrieve highly informative microblogs.

In conclusion, from the results in Table 4, 5, and 6, a mi-
microblog search system should use the concept-based temporal PRF method when searching topically and highly informative relevant documents instead of the word and concept-based lexical PRF methods.

4.5 Additional Experiments

In the remainder of this section, we present further analyses of the various aspects of the proposed wTRM and cTRM methods.

Comparison to existing temporal PRF methods In Sec. 4.4, we compared the proposed temporal PRF methods (wTRM and cTRM) to lexical ones (wRM and cRM). The experimental results show the effectiveness of temporal PRF methods comparing to lexical ones. In this section, we compare the performance of the wTRM and cTRM retrieval methods to the performance of three time-based PRF methods employing the word weighting scheme. The first method, proposed by Li and Croft (2003), incorporates recency into the relevance model of the document prior. The second method, proposed by Keikha et al. (2011), automatically detects this topic-related time for incorporating the temporal property into language modeling frameworks. The third method, proposed by Miyanishi et al. (2013b), combines query-dependent lexical information and document-dependent temporal information of microblogs for word weighting. For comparison, we used the search results reported by Miyanishi et al. (2013b). We briefly compare their performance to wTRM and cTRM because the reported results of the comparative temporal PRF methods were optimized for best performance of Precision at top 30 measure in their paper. Table 7 presents a comparison between our proposed methods and three existing methods. Table 7 shows that wTRM is the best-performing method in both measures for allrel queries. Furthermore, cTRM outperformed other methods in all evaluation metrics for highrel queries. In particular, the difference in AP, and bpref for highrel was statistically significant. For all methods, similar queries and document processing were applied. Similar baselines were reported. Therefore, our novel PRF methods, which extended a language modeling approach from temporal perspective, are effective for microblog searches even when compared to other state-of-the-art temporal PRF methods. Moreover, Table 7 shows that wTRM outperformed cTRM in both measures for allrel queries while cTRM outperformed wTRM in both measures for highrel queries. Nevertheless, none of these differences was statistically significant. In summary, these results also show that concept frequencies over time are important for PRF and the concept-based PRF cTRM is an effective method to retrieve highly relevant documents.

Number of expansion concepts In Sec. 4.4, we tuned the number of concepts k for query expansion using training data. In this section, we assess the effect of increasing the number of expansion concepts. We are particularly interested in addressing the question of whether temporal PRF methods (i.e., wTRM and cTRM) outperformed lexical ones across several k values. Fig. 4 demonstrates that wTRM outperformed wRM, and that cTRM also outperformed cRM across several k values, which reflects that temporal information improves retrieval performance even when using many concepts for query expansion.

Sensitivity to a temporal smoothing parameter In Sec. 4.4, we let temporal smoothing parameter µt = 150 for allrel and µt = 350 for highrel. In this section, we assess how we should smooth language model associated with temporal information. Fig. 5 shows that temporal methods wTRM and cTRM outperform atemporal methods wRM and cRM over allrel and highrel queries across several µt values. In addition, for allrel queries, wTRM outperformed wRM as well as cTRM across several µt values. However, for highrel queries, cTRM outperformed cRM as well as wTRM in almost all µt values. The MAP values of wTRM and cTRM were actually affected by the value of µt, which suggests that the temporal smoothing parameter µt requires different tuning to achieve the best performance for allrel and highrel query sets.

Number of pseudo-relevant documents for temporal evidence In this section, we describe our study of the effect
of increasing the number of feedback documents for temporal information. The large number of feedback documents $N$ means tracking concept’s frequency over the long term. Fig. 6 demonstrates that $wTRM$ and $cTRM$ respectively outperformed $wRM$ and $cRM$ across different feedback documents. However, their performance decreased slightly for $allrel$ and substantially decreased for $highrel$, which indicates that our temporal PRF methods require few feedback documents for concept importance weighting but rather topic-related document for estimating the topically relevant time.

**Expanded concepts** In this section, we present illustrative examples of the types of concepts generated using our model. Figs. 7 and 8 show the top 12 expanded concepts inferred from four PRF methods ($wRM$, $wTRM$, $cRM$, and $cTRM$), respectively, for topics numbered MB109 and MB108. The expanded concepts were ordered by the score of each PRF method. Right panels in Figs. 7 and 8 show the temporal variations of each topic. 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. High density indicates the period during which the topic was described actively. The solid line (Rel) shows the estimate for relevant documents. The dotted line (LM) show the estimate of top 30 retrieved documents by LM with only language filtering, which were used for temporal PRF methods.

In fact, Figs. 7 and 8 clarify that estimating accurate temporal variation of a given topic using temporal PRF methods $wTRM$ and $cTRM$ suggests more topic-related words and concepts than $wRM$ and $cRM$ using only lexical information for their feedback. For example, $wTRM$ and $cTRM$ improved the retrieval performance in AP (0.4454 to 0.5109 and 0.4014 to 0.5843) versus $wRM$ and $cRM$, respectively, because $wTRM$ and $cTRM$ can rank topic-related words and concepts (e.g., $film$, $documentary$, and $oscar nomination$ in MB109) at the top. However, $wTRM$ and $cTRM$ could not find topic-related words and concepts (e.g., $scammed$, $cost$, and $theft cost$ in MB108) and decreased AP values (0.3552 to 0.2185 and 0.3753 to 0.2038) versus $wRM$ and $cRM$, respectively. These results suggest that estimating the relevant time for each topic is important to weight important concepts accurately.

**5 Conclusion**

This paper presented a concept-based query expansion method based on a temporal pseudo-relevance feedback (PRF) model. Unlike existing retrieval models that use only lexical information of concepts, the proposed model effectively combines lexical and temporal properties by modeling temporal variations of concepts in microblogging services. Our empirical results on the Tweets2011 corpus used in

---

7*Gasland* is a documentary movie which has earned an Academy Award nomination for best documentary in 2011.

8The article titled “How Much Does Identity Theft Cost?” was described by many people in Twitter around January 29, 2011.
TREC 2011 and 2012 microblog track demonstrate that incorporating temporal information of concepts into the query expansion method improved retrieval performance significantly. We demonstrated that using multi-term concepts for the temporal PRF method can be useful for retrieving highly relevant documents. Furthermore, our method significantly outperformed existing temporal PRF methods.

Although our concept-based temporal PRF method is effective for microblog search, our temporal PRF method sometimes failed to outperform the lexical one when pseudo-relevant documents failed to estimate topically relevant time. In future work, we plan to incorporate our time-aware latent concept expansion methods into the two-stage relevance feedback framework which can estimate more accurate topically relevant time (Miyanishi, Seki, and Uehara 2013b) in order to further improve retrieval performance.

6 Acknowledgments

This work is partially supported by JSPS KAKENHI Grant Numbers 12J02449 and 25330363.

References

Ounis, I.; Macdonald, C.; Lin, J.; and Soboroff, I. 2011. Overview of the TREC-2011 microblog track. In TREC.
Soboroff, I.; Ounis, I.; and Lin, J. 2012. Overview of the TREC-2012 microblog track. In TREC.