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).

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).

wRM	cTRM (Lexical)	cTRM (Temporal)
jay	jay	carney
carney	carney	jay
qantas	qantas	press secretary
new	new spokesman	jay carney
obama	new	biden spokesman

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 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 ranking mechanism integrating temporal properties into a language model, thereby identifying the important periods for a given topic. Keikha et al. (2011) proposed a time-based relevance model for improving blog retrieval. Moreover, Lin and Efron (2013) reported that a temporal IR method for detecting topically related time significantly improves the microblog search performance. Using the notion of temporal profile (Jones and Diaz 2007), represented as a timeline for a set of documents returned by a search engine, Miyanishi et al. (2013a) proposed the query expansion method, which combines recency and temporal variation in response to a query-dependent temporal property. Efron et al. (2012) proposed document expansion combining lexical and temporal information based on the notion of cluster IR. Miyanishi et al. (2013b) proposed a two-stage relevance feedback approach which conducts PRF method integrating lexical and temporal evidence into its relevance model after relevance feedback with manual microblog document selection. Nevertheless, these existing methods mainly use word information and do not use multi-term concepts even though such concepts can discriminate between relevant and non-relevant documents better. In contrast, our method combines lexical and temporal information of concepts for query expansion by modeling the temporal variation of concepts.

2.2 Concept-Based Information Retrieval

Many researchers have reported recently that the concept-based IR method outperformed the word-based one across many tasks. Most successful works weight concept importance using a Markov Random Field (MRF), which generalizes uni-gram, bi-gram, and other various dependence models. The MRF models have improved retrieval performance significantly, especially for web search, where relevance at high ranks is particularly critical. For example, Metzler and Croft (2005) proposed a query expansion method using the MRF model, which represents term-dependency for multiple terms (i.e., concepts) in a query. Moreover, they combine term dependence with query expansion using the MRF model, called LCE (Metzler and Croft 2007). In fact, LCE outperformed a standard query expansion technique based on a bag-of-words model across several TREC datasets without decreasing search performance with regard to many queries. However, LCE mainly uses the concept frequency on the importance of a query concept. It uses no concept information related to external sources. To overcome these shortcomings, Bendersky et al. (2010) proposed a learning-to-rank approach for concept weighting, which uses internal and external sources, such as Wikipedia, and a query log to obtain concept statistics. In addition, Bendersky et al. (2011; 2012) proposed learning-to-rank frameworks that weight concepts extracted from top retrieved documents by LCE as well as concepts in a query. Moreover, Bendersky and Croft (2012) proposed the query formulation method which uses a combination of concepts represented by hyper-graphs generalizing term-dependencies. On both standard newswire and Web TREC corpora, these concept-importance weighting approaches consistently and significantly outperform widely various state-of-the-art retrieval models. However, these concept weighting approaches do not take account of temporal factors which, as described previously, are important factors for microblog searches.

In contrast to previously reported approaches, this research is mainly motivated by the need for retrieving microblogs leveraging temporal information of concepts. The novelty of our work compared to this previous research is that, for refining an input query, we detect topic-related important concepts that have been frequently described by many microblog users at a specified time period. Developing such an approach is our goal for the present study.

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 (Ponte 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 a i-th query term q_i under the word distribution for document D. The maximum likelihood estimator of $P(q|D)$ 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.

$$P(w|D) = \frac{|D|}{|D| + \mu} P_{ml}(w|D) + \frac{\mu}{|D| + \mu} P(w|C), \quad (1)$$

where μ 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, $P(w|\mathcal{R})$, 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 $\mathcal{R}$. That relevance model weights words w according to the following.

$$\begin{aligned} P(w|\mathcal{R}) &\approx P(w|Q) = \sum_{D \in \mathcal{R}} P(w, D|Q) \\ &= \frac{1}{\mathcal{Z}} \sum_{D \in \mathcal{R}} P(D)P(w, Q|D) \\ &\propto \sum_{D \in \mathcal{R}} P(D)P(w|D) \prod_{i}^{n} P(q_i|D), \quad (2) \end{aligned}$$

where $\mathcal{R}$ is a set of relevant or pseudo-relevant document for query Q and where $\mathcal{Z} = \sum_{w \in \mathcal{V}} \sum_{D \in \mathcal{R}} P(w, D, Q)$ is a normalization factor. When using the top M retrieved documents by the query Q for $\mathcal{R}$, 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 $\mathcal{R}$ (top M retrieved documents) as follows:

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

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 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 $\mathcal{R}$, 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

$$S_{cRM}(c,Q) \propto \sum_{D \in \mathcal{R}} P(D)P(c|D) \prod_{i}^{m} P(\hat{q}_i|D), \quad (4)$$

where $\hat{q}_i$ is a 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 can not 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 by 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 $P(c|\mathcal{R})$ derives from both lexical and temporal information sources. Therefore, we have

$$\begin{aligned} P(c|Q) &= \sum_{D_l \in \mathcal{R}_l} \sum_{D_t \in \mathcal{R}_t} P(c, D_l, D_t|Q) \\ &= \sum_{D_l \in \mathcal{R}_l} \sum_{D_t \in \mathcal{R}_t} P(D_l|c, D_t, Q)P(c, D_t|Q), \quad (5) \end{aligned}$$

where D_l denotes a document from pseudo-relevant documents $\mathcal{R}_l$ and D_t denotes each time (a day in our case) in $\mathcal{R}_t$. Then, as with the work by Efron and Golovchinsky (2011), we apply the simple assumption that the temporal information D_t is independent of the lexical information D_l, so that D_t is dropped from the conditional probability in Eq. 5. Therefore, we have

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

Then, following the notion of bag-of-concepts, we assume that query concepts $\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_m$ and concept c for query

[1]Because the concept frequencies contribute little to the significant improvements in retrieval performance (Macdonald and Ounis 2010), we set $\phi_3(c,C) = 0$.

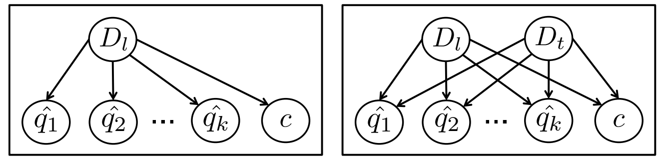

Figure 1: Graphical model representations of concept-based relevance modelling (left) and the proposed concept-based temporal relevance modelling (right).

expansion are sampled identically and independently from a lexical distribution of pseudo-relevant documents, $\mathcal{R}_l$, and a time distribution of ones, $\mathcal{R}_t$ (top N retrieved documents). We have

$$\begin{aligned} P(c|Q) \propto \frac{1}{P(c|Q)} & \sum_{D_l \in \mathcal{R}_l} P(D_l)P(c|D_l) \prod_{j}^{m} P(\hat{q}_j|D_l) \cdot \\ & \sum_{D_t \in \mathcal{R}_t} P(D_t)P(c|D_t) \prod_{j}^{m} P(\hat{q}_j|D_t) \end{aligned}$$

where $P(c|D_l)$ and $P(\hat{q}|D_l)$ denote the probability of concept occurrence in document D; $P(c|D_t)$ and $P(\hat{q}|D_t)$ 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

$$\begin{aligned} S_{cTRM}(c,Q) \overset{rank}{=} & \Bigg\{ \underbrace{\sum_{D_l \in \mathcal{R}_l} P(D_l)P(c|D_l) \prod_{i}^{m} P(\hat{q}_i|D_l)}_{\textbf{Lexical}} \cdot \\ & \underbrace{\sum_{D_t \in \mathcal{R}_t} P(D_t)P(c|D_t) \prod_{i}^{m} P(\hat{q}_i|D_t)}_{\textbf{Temporal}} \Bigg\}^{1/2}, \end{aligned} \quad (6)$$

Here $P(D_l)$ and $P(D_t)$ are uniform over all the distributions in D_l and D_t. The value of $P(c|D_t) \prod_{j}^{|m|} P(\hat{q}_j|D_t)$ 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 $P(c|D_t)$ derived from document time-stamps of pseudo-relevant documents $\mathcal{R}_t$, this PRF model represents real-time feature of a given topic in microblogging services. In addition, because $P(c|D_l) \prod_{i}^{m} P(\hat{q}_i|D_l)$ 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 $P(c|D_t)$, we also use Dirichlet smoothing as with the standard query likelihood model in Eq. 1 because the value of query likelihood $\prod_{i}^{m} P(\hat{q}_i|D_t)$ becomes 0 when a query concept $\hat{q}_i$ does not appear over time in $\mathcal{R}_t$. We have

$$P(c|D_t) = \frac{|D_t|}{|D_t| + \mu_t} \hat{P}_{ml}(c|D_t) + \frac{\mu_t}{|D_t| + \mu_t} P(c|C), \quad (7)$$

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

Name	Type	#Topics	Topic Numbers
TREC 2011	*allrel*	49	1-49
	highrel	33	1, 10-30, 32, 36-38, 40-42, 44-46, 49
TREC 2012	*allrel*	59	51-75, 77-110
	highrel	56	51, 52, 54-68, 70-75, 77-104, 106-110

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, μ_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[2]). 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, $\langle num \rangle$ is a topic number, $\langle title \rangle$ is a user query, and $\langle querytime \rangle$ is the query-time when the query was issued. In our experiments, we use $\langle title \rangle$ 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.

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

`<num>`	MB001
`<title>`	BBC World Service staff cuts
`<querytime>`	Tue Feb 08 12:30:27 +0000 2011

Figure 2: Example topic from the TREC microblog track.

Table 3: Summary of evaluated retrieval methods.

Method	Lexical	Temporal	Concept
wRM	✓		
cRM	✓		✓
wTRM	✓	✓	
cTRM	✓	✓	✓

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 μ_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 $P(c|C) \approx df(c)/N$, 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 μ_t. The sensitivity of a parameter μ_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.

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

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

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

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 adjoiningly 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, $\mathcal{R}_t$, 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 $\mathcal{R}_t$.

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 re-weighting. 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 $S_{cTRM}(c,Q)$. We use the normalized score for concept weighting. For example, the weight of i-th concept is $c_i = \frac{S_{cTRM}(c_i,Q)}{\sum_j^k S_{cTRM}(c_j,Q)}$ when using cTRM. Finally, we combined the expanded concepts of PRF with their weight and the original query as an expanded query. 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 the 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 α, β, and γ respectively denote statistically significant improvements over LM, wRM, and wTRM. The best result per column is marked by boldface.

	allrel			*highrel*		
Method	**AP**	**Rprec**	***bpref***	**AP**	**Rprec**	***bpref***
LM	0.2936	0.3313	0.3103	0.2130	0.2286	0.1933
wRM	0.3502^{α}	0.3868^{α}	0.3594^{α}	0.2473^{α}	0.2537	0.2242
wTRM	$\mathbf{0.3726^{\alpha}_{\beta}}$	$\mathbf{0.4089^{\alpha}}$	$\mathbf{0.3872^{\alpha}_{\beta}}$	$\mathbf{0.2580^{\alpha}}$	$\mathbf{0.2705^{\alpha}}$	$\mathbf{0.2361^{\alpha}}$

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

	allrel			*highrel*		
Method	**AP**	**Rprec**	***bpref***	**AP**	**Rprec**	***bpref***
LM	0.2936	0.3313	0.3103	0.2130	0.2286	0.1933
cRM	0.3385^{α}	0.3725^{α}	0.3479^{α}	0.2511^{α}	0.2696^{α}	0.2356^{α}
cTRM	$\mathbf{0.3644^{\alpha}}$	$\mathbf{0.4058^{\alpha}_{\beta}}$	$\mathbf{0.3825^{\alpha}_{\beta}}$	$\mathbf{0.2694^{\alpha}_{\beta}}$	$\mathbf{0.2770^{\alpha}}$	$\mathbf{0.2527^{\alpha}}$

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[6] 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 topically 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 results 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.

[6]http://www.mansci.uwaterloo.ca/~msmucker/software/paired-randomization-test-v2.pl

Table 6: Performance comparison of the standard word-based PRF method and the proposed concept-based temporal one. Superscripts α and β respectively denote statistically significant improvement over wRM, and cTRM. Best result per column is marked by boldface.

	allrel			*highrel*		
Method	**AP**	**Rprec**	***bpref***	**AP**	**Rprec**	***bpref***
wRM	0.3502	0.3868	0.3594	0.2473	0.2537	0.2242
cTRM	**0.3644**	**0.4058**	**0.3825**	$\mathbf{0.2694^{\alpha}}$	**0.2770**	$\mathbf{0.2527^{\alpha}}$

Table 7: Performance comparison of the existing temporal PRF methods and the proposed temporal ones. Statistically significant difference of wTRM and cTRM over the baselines are marked using α, β and γ, for EXRM (Li and Croft 2003), TBRM (Keikha, Gerani, and Crestani 2011), and QDRM (Miyanishi, Seki, and Uehara 2013b) baselines, respectively. Best result per column is marked by boldface.

	allrel			*highrel*		
Method	**AP**	**Rprec**	***bpref***	**AP**	**Rprec**	***bpref***
EXRM	0.3560	0.3846	0.3634	0.2433	0.2485	0.2202
TBRM	0.3539	0.3862	0.3607	0.2347	0.2384	0.2071
QDRM	0.3568	0.3829	0.3642	0.2522	0.2622	0.2306
wTRM	**0.3726**	**0.4089**	**0.3872**	0.2580	0.2705^{β}	0.2361
cTRM	0.3644	0.4058	0.3825	$\mathbf{0.2694^{\alpha}_{\beta}}$	**0.2770**	$\mathbf{0.2527^{\alpha}_{\beta}}$

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 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 shows 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 (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

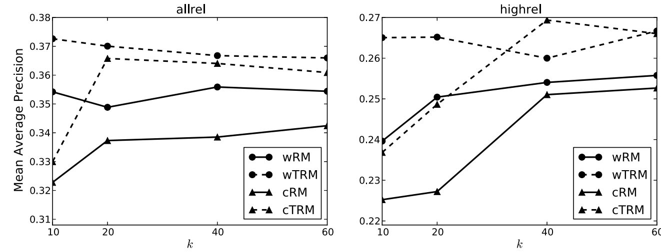

Figure 4: Effects of increasing the number of expansion concepts k on the retrieval effectiveness of the *allrel* and *highrel*queries. The x-axis shows parameter k. The y-axis shows the values in MAP.

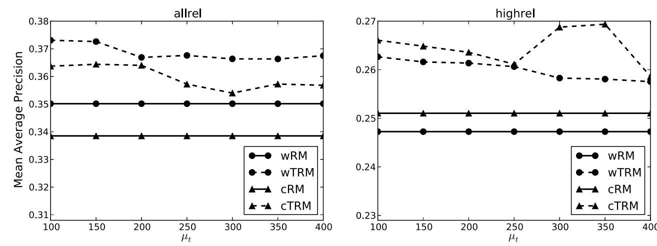

Figure 5: Sensitivity to a temporal smoothing parameter μ_t on the retrieval effectiveness of the *allrel* and *highrel* queries. The x-axis shows parameter k. The y-axis shows values in MAP.

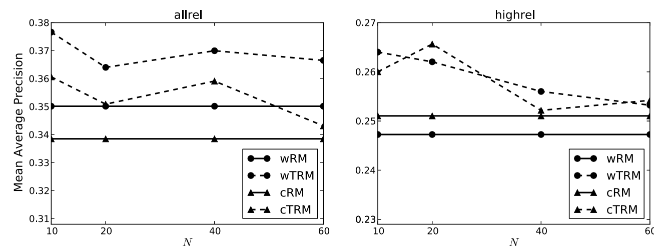

Figure 6: Effect of increasing the number of feedback documents for temporal information on the retrieval effectiveness of the *allrel* and *highrel* queries. The x-axis shows parameter k. The y-axis shows values in MAP.

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 $\mu_t = 150$ for *allrel* and $\mu_t = 350$ 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 μ_t values. In addition, for *allrel* queries, wTRM outperformed wRM as well as cTRM across several μ_t values. However, for *highrel* queries, cTRM outperformed cRM as well as wTRM in almost all μ_t values. The MAP values of wTRM and cTRM were actually affected by the value of μ_t, which suggests that the temporal smoothing parameter μ_t 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

wRM	wTRM	cRM	cTRM
oscar	oscar	truth	oscar
industry	industry	truth gasland	industry
truth	truth	oscar	truth
nod	nod	industry	nod
fundamentally	nomination	gasland fundamentally	oscar nod
dishonest	film	dishonest	oscar nomination
moore	documentary	fundamentally dishonest	nomination
gore	moore	fundamentally	truth gasland
nomination	gore	gasland moore	film
news	news	gore	gasland moore
receives	filmmakers	more	moore gore
boos	boos	nod	gore

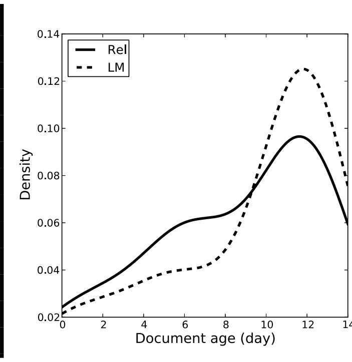

Figure 7: Twelve most likely one or two word concepts discovered by wRM, wTRM, cRM, and cTRM for the query "*Gasland*" (MB109), showing improved results with temporal PRF methods wTRM and cTRM. Left figure shows temporal variations of a topic numbered MB109.

wRM	wTRM	cRM	cTRM
best	best	best identity	advocate
advocate	advocate	advocate	best
cost	ring	best	advocate best
www	alleged	advocate best	best identity
hub	www	theft cost	alleged identity
restoration	hub	cost	theft ring
spears	restoration	restoration www	ring
prepaidlegal	spears	com hub	alleged
com	prepaidlegal	hub spears	com hub
colorado	com	protection restoration	hub spears
experts	colorado	www	protection restoration
scammed	experts	prepaidlegal com	spears

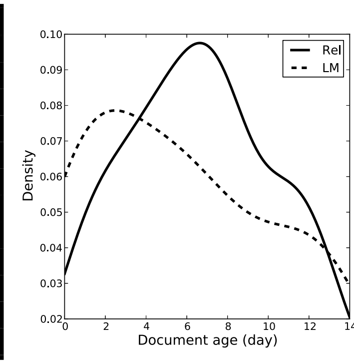

Figure 8: Twelve most likely one or two word concepts discovered by wRM, wTRM, cRM, and cTRM for the query "*identity theft protection*" (MB108), showing harmed results with temporal PRF methods wTRM and cTRM. Left figure shows temporal variations of a topic numbered MB108.

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[7]) at the top. However, wTRM and cTRM could not find topic-related words and concepts (e.g., *scammed*, *cost*, and *theft cost* in MB108[8]) 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.

[8]The 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

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

Bendersky, M., and Croft, W. B. 2012. Modeling higher-order term dependencies in information retrieval using query hypergraphs. In *SIGIR*, 941–950.

Bendersky, M.; Metzler, D.; and Croft, W. B. 2010. Learning concept importance using a weighted dependence model. In *WSDM*, 31–40.

Bendersky, M.; Metzler, D.; and Croft, W. B. 2011. Parameterized concept weighting in verbose queries. In *SIGIR*, 605–614.

Bendersky, M.; Metzler, D.; and Croft, W. B. 2012. Effective query formulation with multiple information sources. In *WSDM*, 443–452.

Buckley, C., and Voorhees, E. M. 2000. Evaluating evaluation measure stability. In *SIGIR*, 33–40.

Buckley, C., and Voorhees, E. M. 2004. Retrieval evaluation with incomplete information. In *SIGIR*, 25–32.

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

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

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

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

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

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

Krovetz, R. 1993. Viewing morphology as an inference process. In *SIGIR*, 191–202.

Lang, H.; Metzler, D.; Wang, B.; and Li, J.-T. 2010. Improved latent concept expansion using hierarchical Markov random fields. In *CIKM*, 249–258.

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

Lease, M. 2009. An improved Markov random field model for supporting verbose queries. In *SIGIR*, 476–483.

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

Lin, J., and Efron, M. 2013. Temporal relevance profiles for tweet search. In *TAIA*.

Macdonald, C., and Ounis, I. 2010. Global statistics in proximity weighting models. In *SIGIR Web N-gram Workshop*.

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

Metzler, D., and Cai, C. 2011. USC/ISI at TREC 2011: Microblog track. In *TREC*.

Metzler, D., and Croft, W. B. 2005. A Markov random field model for term dependencies. In *SIGIR*, 472–479.

Metzler, D., and Croft, W. B. 2007. Latent concept expansion using Markov random fields. In *SIGIR*, 311–318.

Metzler, D.; Cai, C.; and Hovy, E. 2012. Structured event retrieval over microblog archives. In *NAACL-HLT*, 646–655.

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

Miyanishi, T.; Seki, K.; and Uehara, K. 2013b. Improving pseudo-relevance feedback via tweet selection. In *CIKM*, 439–448.

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

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

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

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW*, 851–860.

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

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

Strohman, T.; Metzler, D.; Turtle, H.; and Croft, W. 2005. Indri: a language model-based search engine for complex queries. In *ICIA*, 2–6.

Teevan, J.; Ramage, D.; and Morris, M. 2011. #TwitterSearch: a comparison of microblog search and web search. In *WSDM*, 35–44.

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