Modeling Temporal Evidence from External Collections

Flávio Martins
NOVA LINCS
School of Science and Technology
Universidade NOVA de Lisboa
Caparica, Portugal
flaviomartins@acm.org

João Magalhães
NOVA LINCS
School of Science and Technology
Universidade NOVA de Lisboa
Caparica, Portugal
jm.magalhaes@fct.unl.pt

Jamie Callan
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
callan@cs.cmu.edu

ABSTRACT

Newsworthy events are broadcast through multiple mediums and prompt the crowds to produce comments on social media. In this paper, we propose to leverage on this behavioral dynamics to estimate the most relevant time periods for an event (i.e., query). Recent advances have shown how to improve the estimation of the temporal relevance of such topics. In this approach, we build on two major novelties. First, we mine temporal evidences from *hundreds of external sources* into topic-based external collections to improve the robustness of the detection of relevant time periods. Second, we propose a formal retrieval model that *generalizes the use of the temporal dimension* across different aspects of the retrieval process. In particular, we show that temporal evidence of external collections can be used to (i) infer a topic's temporal relevance, (ii) select the query expansion terms, and (iii) re-rank the final results for improved precision. Experiments with TREC Microblog collections show that the proposed time-aware retrieval model makes an effective and extensive use of the temporal dimension to improve search results over the most recent temporal models. Interestingly, we observe a strong correlation between precision and the temporal distribution of retrieved and relevant documents.

KEYWORDS

Microblog search, social media, learning to rank, time-aware ranking models, temporal information retrieval

1 INTRODUCTION

A networked world and the increasing pervasiveness of Internet access enables the rapid adoption of new online communication mediums to discuss current events. Previous research has explored this symbiosis between Twitter and the news [17, 29] and linked

the two mediums [12, 33]. Events are discussed on the Web as they happen and people following them can add to the conversation immediately. Hence, improving the *temporal relevance estimation* for searching such events became a significant research priority.

Nowadays, the state-of-the-art Web search systems are based on learning to rank feature models that combine multiple text retrieval functions as well as other features. Relevance on Twitter has many dimensions: authority, popularity, freshness, geographical context, and topical relevance. Previously, time-aware ranking research explored the assumption that fresh documents are more relevant [19]. Later models revised this assumption in line with what is observed in Twitter: for time-sensitive queries, documents tend to cluster temporally [9, 11]. Our approach is based on the intuition that discussions about a topic and its subtopics are likely to occur around the same time across multiple mediums.

The rationale is that newsworthy events trigger a cascade of activity on the Web and Twitter. This information can be useful for ranking and, in some cases, can be gathered with ease. The news often have a good coverage of current topics, clean journalistic language, and reliable timestamps. Thus, it is desirable to mine news sources to offer more context to the *tweets* as well as to the users' queries intent. In particular, we aim to explore the *crowd aggregation effect* to extract temporal evidence from news verticals. Temporal evidence is further used to refine the selection of query expansion terms and to estimate query topics temporal relevance. This approach is completed with the re-ranking of the final search results leading to improved precision. Hence, the proposed method brings a series of novel contributions:

- Explore the crowd effect by aggregating posts published by news sources into topic-based external collections;
- Mining of crowds' temporal evidence at different granularities (i.e., *verticals*, *documents*, and *terms*);
- A formal time-aware ranking model that unifies multiple temporal features into a single comprehensive retrieval model.

Including the temporal dimension at the different steps of the search engine pipeline, improves the accuracy of several retrieval tasks, leading to greater overall gains. This is possible because, the temporal dimension introduces stronger evidence in many decision tasks (e.g., selection of query expansion terms). Evaluation on the TREC 2013 and TREC 2014 Microblog Track datasets shows that the proposed retrieval model outperforms state-of-the-art methods.

This paper is organized as follows: in Section 2 we present the related work; in Section 3 the formal temporal ranking model is detailed and the following sections detail its implementation; evaluation is presented in Section 5; and a more fine-grained discussion of results in Section 6.

2 RELATED WORK

In the past, several authors have proposed to use multiple collections to improve search results. Bendersky et al. [3] point to the limitation that standard query formulation tasks such as term weighting and query expansion often use a single source of information. In their query expansion experiments, they combined multiple information sources from newswire and web corpora and found better retrieval effectiveness than when using a single source of information. Weerkamp et al. [35] developed a novel query modeling framework to combine evidence from multiple external collections. An interesting property of this model is that, if we assume that the query-dependent collection importance $P(q \mid c)$ is uniformly distributed and that the importance of a document in the collection is $P(d \mid c) = 1/|\mathcal{R}_c|$, we arrive at the formulation of Mixture of Relevance Models (MoRM) proposed by Diaz and Metzler [10].

Several time-based pseudo-relevance feedback methods were proposed for retrieval in time-sensitive collections using the relevance modeling framework. Keikha et al. [15] proposed time-based relevance models where they assume that the publishing date has an effect on the terms. They introduce a generative model of the query that first selects a date and then a term based on the time and query. They found this approach was able to improve the coverage of the expanded query over the different subtopics by using temporal information to weight and select expansion terms. Choi and Croft [4] extended the framework proposed by Keikha et al. [15] by making a simplifying assumption that $P(d \mid T, q)$ can be equal to $P(d \mid q)$ since the temporal dimension is incorporated already in choosing d. In this formulation, a relevance model for each time period is estimated using the retrieved documents published in time the time period T. Each of the relevance models are then weighted by $P(T \mid q)$ to obtain the final expansion terms over all the time periods. Arguello et al. [2] also explored the use of external collections, in the form of verticals, to leverage first-order statistics from verticals to improve search results. The method proposed in this paper also explores multiple collections (i.e., organized into verticals) but moves far beyond first-order statistics.

Recent time-dependent ranking approaches have resorted to learning to rank techniques [6, 14, 24] and temporal query detection [8] that exploit non-temporal and temporal features. Dai et al. [8] propose to run each query against a set of rankers, which are weighted based on the temporal profile of a query, and therefore minimize the risk of degraded performance due to misclassifying the query in terms of recency intent. Metzler et al. [24] defined *microblog event retrieval* as a search task that goes beyond ad hoc retrieval. To uncover subtopics from these streams of very short and noisy posts they proposed a temporal query expansion technique. Their technique divides the timeline into time spans of one hour, and ranks them according to the proportion of messages posted during the time spans that match the query. The *burstiness score* weights terms, so that when counts for the occurrences of terms are higher than usual, these terms will have higher weights.

Microblog retrieval has very specific and time-pressed requirements. For instance, Jones and Diaz [13] note that queries that favor recency are just a subset of the time-sensitive queries. Along this vein of thought, several authors [26, 36] proposed to leverage the temporal distribution of the pseudo-relevant documents. Whiting

et al. [36] combines pseudo-relevant document term distribution and temporal collection evidence using a variant of PageRank over a weighted graph that models the temporal correlation between n-grams. Another approach by Peetz et al. [26] leverages the temporal distribution of the pseudo-relevant documents themselves. For each query, bursty time periods are identified and then documents from these periods are selected for feedback. The query model is updated with new terms sampled from the higher quality documents selected. Dakka et al. [9] evaluated a time-sensitive pseudo-relevance feedback method on manually selected topic subsets from various TREC collections (e.g., TREC News Archive, TREC Time-sensitive Queries). Dakka et al. [9] identified the need to find the important time periods for time-sensitive queries and to integrate temporal relevance in the ranking model. Their ranking model explicitly splits the lexical and temporal evidence in the documents: w_d, the words in the document and t_d, the document's timestamp: $P(d \mid q) \propto P(w_d \mid q) \cdot P(t_d \mid q)$. They propose techniques to estimate the $P(t_d \mid q)$ using histograms, however this method might be cumbersome as the calculation of histogram bins is linked to many parameters.

Recently, modeling temporal relevance was shown to be effective for searching time-sensitive collections. Craveiro et al. [7] explored the segmentation of textual news articles, so that it can be leveraged in the query expansion process to focus the expansion terms temporally. Efron et al. [11] proposed a general and principled retrieval model for microblog search with temporal feedback. Their approach models the temporal density of a query $P(t \mid q)$ with a kernel density estimate, with all the advantages brought by this method: the natural smoothness of the resulting function and a fully automated way to estimate the model variables (e.g., bandwidth selection is data-driven, a function of the initial rank). This estimated temporal relevance is then employed to re-rank documents with a log-linear model. Martins et al. [22] achieved state-of-the-art results when using multiple external sources such as Wikipedia edits, views and newswire articles. Following the same rationale, that term expansions should be biased to draw from documents from relevant (bursty) time periods. Rao and Lin [27] proposed capturing these by estimating the parameters of a continuous hidden Markov model that best explains the sequential dependencies in the temporal distribution of documents retrieved in the initial feedback step, computing the most likely state sequence using the Viterbi algorithm, and drawing terms only from bursty states. These findings have found their way into the architecture of search indexes. Wang and Lin [34] examined how the main index collection could be partitioned (i.e., by day, by source, etc.). This provides the important insight that collections can actually be partitioned over time.

In contrast to previous work, we propose to use multiple news verticals to robustly identify the relevant time periods for each query, instead of relying only on the temporal distribution of pseudo-relevant documents [9] or first-order statistics from verticals [2].

3 MODELING TEMPORAL EVIDENCE FROM EXTERNAL COLLECTIONS

Consider a retrieval corpus containing N documents, represented by D. To integrate the temporal relevance component in the ranking model Dakka et al. [9] decomposed the document in two different

parts: lexical evidence, the words in the document (w_d), and temporal evidence, the document's timestamp (t_d). We consider an augmented ranking model that contemplates query-independent signals or metadata from the document, m_d, in addition to the lexical and temporal evidence as follows:

$$
\begin{aligned}
P(d \mid q) &= P(w_d, t_d, m_d \mid q) \\
&\propto P(w_d \mid q) \cdot P(t_d \mid q) \cdot P(m_d \mid q) \\
&\propto \underbrace{P(q \mid w_d) \cdot P(w_d)}_{\text{query-likelihood model}} \cdot P(t_d \mid q) \cdot P(m_d)
\end{aligned}
\tag{1}
$$

where the final formulation follows from the two following steps: First, by applying the Bayes' rule to $P(d \mid q)$ and eliminating the quotient $P(q)$ based on the rank equivalence to get the well-known query-likelihood retrieval model. Second, by assuming the independence between document metadata and the query, $P(m_d)$ can be taken as the query-independent importance of the document.

To instantiate the ranking model from Eq. (1), we need to estimate three components: lexical, temporal, and query-independent. The lexical component can be estimated using *relevance models* (see Section 3.2) or standard query-likelihood, where we assume that $P(w_d)$ is uniform. In this paper, we focus on estimating the temporal component using external collections (see Section 3.1). The query-independent component can be estimated using values extracted from the metadata of the document (see Table 1).

To estimate the temporal component, former models [9, 11] assume that relevant temporal information is only available from the search corpus itself, D, for instance via the temporal distribution of an initial set of feedback documents. However, temporal feedback on the corpus alone can be boosted by external sources [22]. Therefore, we propose improving the estimation of temporal relevance using external collections, in addition to the retrieval corpus:

$$
\begin{aligned}
P(t_d \mid q) &= P(t_d \mid q, D, C) \\
&= P(t_d \mid q, D) \cdot P(t_d \mid q, C),
\end{aligned}
\tag{2}
$$

where the last step follows if we assume that temporal evidence can be extracted from the search corpus D and from the external collections $C = \{c_1 c_2 \cdots c_{|C|}\}$ independently. The first part can be estimated from the temporal distribution of feedback documents retrieved using the query q. We calculate the temporal relevance according to the external collections as described in Section 3.1.

We also propose to generate query expansions to improve the document ranking (i.e., the lexical component) by leveraging the external collections to estimate time-based *relevance models*. In Section 3.2, we present a novel external time-based relevance model to generate expanded query models for retrieval in the corpus. The expanded query model is computed by taking into account lexical as well as temporal evidence contained in the external collections.

3.1 External Temporal Relevance

For a given query, different collections yield different temporal relevance estimates (i.e., different probability distributions of relevance over time). Therefore, we need to extend Eq. (2) to combine all the different temporal relevance estimates from each external collection into a single robust estimate. In our approach, we combine them

using a weighted mixture of probability distributions

$$
\begin{aligned}
P(t_d \mid q, C) &\propto \sum_{c \in C} P(t_d \mid q, c) \cdot P(c \mid q) \\
&\propto \sum_{c \in C} P(t_d \mid q, c) \cdot P(q \mid c) \cdot P(c),
\end{aligned}
\tag{3}
$$

where $P(t_d \mid q, c)$ is the importance of time t_d for the query q in the collection c, $P(q \mid c)$ is the relevance of the collection c to the query q, and $P(c)$ is the query-independent collection prior.

Considering that we may have many external collections, the calculation of temporal relevance over all of them raises efficiency concerns. To solve this problem we follow *federated search* research [23, 31], and consider that only a few collections contain most of the temporal evidence for a given query q. Therefore, we can use only those collections to provide an adequate approximation

$$
P(t_d \mid q, C) \propto \sum_{c \in C_q} P(t_d \mid q, c) \cdot P(q \mid c) \cdot P(c),
\tag{4}
$$

where C_q is a ranking of the most relevant collections to query q, and the query-independent prior of the collection is considered uniform $P(c) = 1/|C_q|$

To estimate the relevance of each collection c for a query q, represented by $P(q \mid c)$, we consider a similar approach to the ReDDE resource selection algorithm [32]. Considering M_k, the final single ranking obtained by merging all the results retrieved from the selected collections C_q, the relevance of collection c is given by the ratio between the number of its documents that make it into the top ranking, M_c, by the total documents retrieved, M_k:

$$
P(q \mid c) = \frac{|M_c|}{|M_k|}
\tag{5}
$$

There are several other options that can be used to estimate $P(q \mid c)$, they include other resource selection algorithms [1, 16, 30, 32, 35].

3.1.1 Vertical Temporal Feedback. For each document d we would like to find $P(t_d \mid q, c)$, the probability of relevance of its timestamp t_d according to vertical c and the query q. This probability follows the joint distribution $f_c(t_d)$,

$$
P(t_d \mid q, c) \sim f_c(t_d).
\tag{6}
$$

Following Efron et al. [11], we estimate the probability density function $f_c(t_d)$ by learning the distribution of feedback documents using a weighted kernel density estimation method:

$$
f_c(t) = \frac{1}{nh} \sum_{d \in \mathcal{R}_c} \lambda_d \, K\left(\frac{t - t_d}{h}\right)
\tag{7}
$$

where t is the timestamp of the input document, $\mathcal{R}_c$ is the set of retrieved documents from the collection c and t_d corresponds to these documents' timestamps. The kernel function $K(z)$ corresponds to the Gaussian kernel $\mathcal{N}(z, 0)$, and the optimal bandwidth can be estimated by a data-driven method such as Silverman's rule-of-thumb $h^* \approxeq 1.06 \, \sigma \, n^{-1/5}$. Finally, λ_d, is a non-negative weight on timestamp t_d, to weight each timestamp by its importance. The weight λ_d of each document's timestamp is based on its relevance to the query, for instance, the document's query-likelihood model retrieval score or estimated from its position in the rank.

3.2 External Time-based Relevance Models

Relevance models provide a framework for term selection and estimation of the importance of terms for query expansion [18]. We propose to estimate relevance models and generate a final query q' using external collections $\mathcal{C}$, leveraging their temporal evidence,

$$P(q \mid w_d, \mathcal{C}) \approx P(q' \mid w_d). \tag{8}$$

Let θ_q be the original query model and θ_{F_C} an estimated feedback query model based on feedback documents $d_1 \cdots d_k$ from multiple external collections. Inspired by Zhai and Lafferty [38], the final query model is $\theta_{q'} = (1 - \alpha)\,\theta_q + \alpha\,\theta_{F_C}$. In this formulation, the final query is a linear combination of the original query model, $P(w \mid \theta_q)$, and the estimated feedback query model, $P(w \mid \theta_{F_C})$, using external collections:

$$P(w \mid \theta_{q'}) = \lambda \cdot P(w \mid \theta_q) + (1 - \lambda) \cdot P(w \mid \theta_{F_C}), \tag{9}$$

where the original query is modeled using its maximum-likelihood estimate $P(w \mid \theta_q) = \#(w, q)/|q|$. Time is introduced in the second parcel of the above expression to improve the estimation of the feedback query expansion terms. To this end, we integrate temporal feedback into term selection to make it time-aware.

We start by estimating the feedback query model by leveraging pseudo-relevant documents from multiple external collections using a formulation proposed by Weerkamp et al. [35]:

$$\begin{aligned} P(w \mid \theta_{F_C}) &\propto \sum_{c \in \mathcal{C}} P(w \mid q, c) \cdot P(c \mid q) \\ &\propto \sum_{c \in \mathcal{C}} P(c \mid q) \sum_{d \in \mathcal{R}_c} P(w \mid d, q) \cdot P(d \mid c) \end{aligned} \tag{10}$$

where we limit the computation of $P(w \mid q, c)$ to the top documents retrieved from each individual collection, $\mathcal{R}_c$. Furthermore, if we consider $P(d \mid c)$ to be uniform (i.e., equal to $1/\mathcal{R}_c$), we obtain the following formulation:

$$\propto \sum_{c \in \mathcal{C}} P(c \mid q) \frac{1}{|\mathcal{R}_c|} \sum_{d \in \mathcal{R}_c} P(w \mid d) \cdot P(q \mid d) \tag{11}$$

In this formulation, term selection is blind to the temporal dimension because time has no influence on the importance of expansion terms. However, in time-sensitive collections, the words in documents published on relevant time periods are more important and therefore should have a higher weight in the final expanded query. Therefore, the formulation above is modified by considering both lexical and temporal components of the documents:

$$\propto \sum_{c \in \mathcal{C}} P(c \mid q) \frac{1}{|\mathcal{R}_c|} \sum_{d \in \mathcal{R}_c} P(w \mid w_d, t_d) \cdot P(q \mid w_d, t_d) \tag{12}$$

$$\approx \sum_{c \in \mathcal{C}_q} P(c \mid q) \frac{1}{|\mathcal{R}_c|} \sum_{d \in \mathcal{R}_c} P(w \mid w_d) \cdot P(q \mid w_d) \cdot P(t_d \mid q) \tag{13}$$

the last step stems from the fact that the probability of word w for document d depends only on the document content w_d and is independent of its timestamp t_d, and that the probability of the query $P(q)$ is constant. Hence, this formulation assumes that $P(t_d \mid q)$ can be estimated from each document timestamp t_d. As in Section 3.1.1, we use kernel density estimation [11] to provide a smooth estimate of $P(t_d \mid q)$.

Finally, an approximate relevance model is calculated using only the most relevant collections $\mathcal{C}_q$ to the query q, since they should contribute the most to the estimation of the final expansion term weights. As in Eq. (5), we assume $P(c)$ to be constant and uniform.

3.3 Estimation over Discrete Time Periods

In the previous section, we proposed a method that assumes a continuous approach to the temporal dimension of relevance, using kernel density estimation to predict the importance of specific points in time. However, it is linked to this specific estimation method while previous methods of estimating temporal relevance exist (e.g., volume-based, histogram-based, window-based) and new methods will be proposed in the future. Therefore, in this section, we discuss a general generative model of the query that relies instead on a discrete partitioning of a timeline into time periods. For each w, it first selects a collection, then a time period, and then a term based on the collection, time period, and the query. Formally,

$$\begin{aligned} P(w \mid \theta_{F_C}) &= \sum_{c \in \mathcal{C}} \sum_T P(w \mid T, q, c) \cdot P(c \mid T, q) \cdot P(T \mid q) \\ &\propto \sum_{c \in \mathcal{C}} P(c \mid q) \sum_T P(w \mid T, q, c) \cdot P(T \mid q), \end{aligned} \tag{14}$$

where $P(w \mid T, q, c)$ is the importance of the word w in time period T (e.g., day, hour) for the query q given collection c, $P(c \mid T, q)$ is the importance of collection c in the time period T for the query q, and $P(T \mid q)$ is the importance of the time period T to the q. The last step follows if we assume the importance of a collection to a query to be independent from any given time period, $P(c \mid T, q) = P(c \mid q)$.

A similar deduction to what was followed in the previous section, leads to a discrete model, which is more compatible with previous research in temporal information retrieval.

4 LEARNING TO RANK MODEL USING EXTERNAL TEMPORAL EVIDENCE

We are now ready to plug-in the temporal evidence and the time-based relevance models from multiple verticals into a common ranking model. To combine the different temporal features extracted from multiple query-specific verticals, we first re-write Eq. (1) as the following log-linear model

$$\log P(d \mid q) \propto Z + \log P(q \mid w_d) + \log P(t_d \mid q) + \log P(m_d), \tag{15}$$

where we can replace $P(t_d \mid q)$ by the temporal relevance over D and $\mathcal{C}$, and the query q by the expanded query q'. Then we can use learning to rank algorithms to learn the optimal weights of the different components using a separate dataset for training, we have

$$\log P(d \mid q) = Z + \sum_i \alpha_i \log P_i(q' \mid w_d) \tag{16}$$

$$+ \beta \log P(t_d \mid q', D) \tag{17}$$

$$+ \gamma \log \sum_{c \in \mathcal{C}_q} P(t_d \mid q, c) \cdot P(q \mid c) \tag{18}$$

$$+ \sum_j \delta_j \log P(m_d^j), \tag{19}$$

where $P(q' \mid w_d)$ is the retrieval score of the document, d, given the expanded query, q'. Instead of using a single estimate of the lexical component, $P(q' \mid w_d)$, more accurate results are obtained

by combining multiple estimates provided by different retrieval models (Eq. (16)). The weights α_i indicate the confidence in each retrieval model's estimate. Since query expansion is used, the documents used for temporal feedback (see Eq. (17)) are retrieved using the expanded query. This additional temporal feedback feature according to the corpus, D, itself, is added on top of the estimation of temporal relevance from the external collections introduced by Eq. (18). To account for different ways to estimate the importance of a document from metadata we introduce Eq. (19). Next, we discuss the relationship between each feature of the above ranking model and the formal model.

Table 1: Learning to rank features.

Feature name	Feature description
Doclen	Document length.
#URL	URL count.
#hashtags	Hashtags count.
#mentions	Mentions count.
hasURL	1 if it contains URL, otherwise 0.
hasHashtags	1 if it contains Hashtags, otherwise 0.
hasMentions	1 if it contains Mentions, otherwise 0.
isReply	1 if it is a Reply, otherwise 0.
#statuses	Total number of posts.
#followers	Total number of followers.

Used in the learning to rank methods, including $KDE+KDE_E+RMT_E$.

Learning to Rank Features. The proposed model is composed of four main components that capture different aspects of search relevance in time-sensitive collections. First, we employ three different retrieval models to obtain textual matching scores, Eq. (16). They are, the query-likelihood retrieval model with Dirichlet prior smoothing (LM.Dir), BM25, and IDF. Second, Eq. (17) includes a temporal feedback feature [11], calculated over the documents retrieved from the main corpus D with the expanded query q'.

Third, the proposed model generalizes the integration of temporal evidence from external collections, Eq. (18), aggregated into a single score. In Eq. (18), the importance of the publishing timestamp of the document t_d according to the external collections is estimated by a summation over the likelihood of each selected verticals. For each vertical, $P(t_d \mid q, c)$ returns the likelihood that an instant represented by t_d is relevant to the query q according to, $\mathcal{R}_c(q)$. The coefficients α_i, β, γ, and δ_j correspond to the feature weights. In contrast to previous work that often relies on a single source of temporal evidence, e.g., corpus, the proposed approach contemplates the use of several external collections. The calculation of the temporal evidence feature over the documents retrieved from query-specific verticals can provide a more robust estimation of the relevant time periods for each query.

Fourth, many non-temporal and query-independent features Eq. (19), were added to improve effectiveness further, such as quality features [5] and other commonly used features in learning to rank approaches to microblog search [37]. Table 1 lists the set of features. This set of features captures microblog-specific information that is useful for ranking such as, number of statuses, number of followers, number of URLs, and number of hashtags. The number of words in the tweet was added as a feature to boost longer documents.
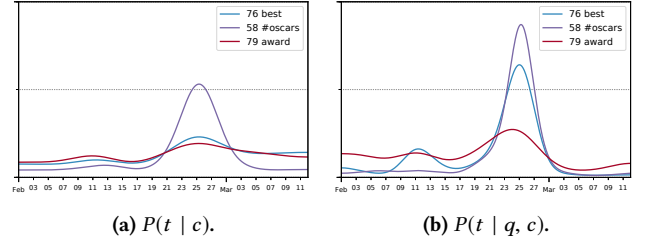

(a) $P(t \mid c)$. **(b)** $P(t \mid q, c)$.

Figure 1: Temporal profiles of queries and collections.

4.1 Example: Temporal Evidence from External Collections

Let's examine an example in light of the described model. Consider the 85th Academy Awards ceremony that took place on February 24 2013, at the Dolby Theatre in Hollywood, Los Angeles. The top award winner, Argo, winning the Oscar Award for *Best Picture*, was a movie starring Ben Affleck. This event sparked multiple processes on the Web, such as the dissemination of news articles about the event, and discussions and commentary on Twitter, Figure 1. A person interested in surveying the general commentary and opinions could procure a list of relevant accounts to monitor posts in real-time. However, journalists and other searchers would most likely use a search engine to find more general information outside one's circle about specific aspects of the event.

We dissected how multiple accounts from news outlets and other verified account on Twitter organized into different topical shards can be used in the search process. Using the TREC Microblog query MB195 - "Argo wins Oscar", we plot two graphs that show the temporal distribution of results at the different stages of the framework. Firstly, in Figure 1a we show the estimation of relevant time periods using only the global statistics of each shard, via kernel density estimation over the timestamps of all of its documents.

In this example we use a resource selection algorithm [1] to select the three most useful shards for the query. All three topical shards selected exhibited a larger probability around the time of the live broadcast. Identified by the word "#oscars", Shard 58, is relatively more bursty than the others. Secondly, since topical shards are too broad we can fine-tune the estimation of the relevant time periods for a given query by finding a further subset of documents that are related to the query. In Figure 1b, we improved the estimation by searching over the same topical shards selected by the resource selection algorithm and for each one using for estimation the documents retrieved by the query "Argo wins Oscar". Shard 79 identified by the word "award" seems to be less focused than the others. The key insight from this comparison is that two topical shards have the most useful temporal information.

5 EXPERIMENTAL METHODOLOGY

This section presents the evaluation of the methods described in the previous sections on the TREC microblog search test-bed. In the TREC Microblog track problem of retrospective ad hoc retrieval, the user wishes to find the most up-to-date and relevant posts. The task can be summarized as: at time t, find *tweets* about topic q. Therefore, systems should favor highly informative *tweets* relevant to the query topic that were published before the query time.

5.1 Protocol

Our experiments delve into the problem of re-ranking *tweets* sampled using a standard retrieval method (i.e., query-likelihood model) taking into account temporal crowd signals from different sources. In our experiments, we follow TREC and report the MAP and P30 results. Statistical significance of effectiveness differences are determined using two-sided paired *t*-tests following Sakai [28].

Filtering Duplicates and Languages. In the collection used, *retweets* are considered not relevant because they are seen as duplicate documents. Therefore, we filtered *Twitter-style retweets* using the tweet metadata available, and we also filter out *RT-style retweets* that start with *RT*. Moreover, assessors evaluated only relevant *tweets* written in English, therefore we use the language filter ldig[1] to remove *tweets* in other languages.

5.2 Datasets

5.2.1 TREC Microblog. The Tweets2013 dataset is the most comprehensive evaluation resource for *ad hoc* retrieval on social media to date. The Tweets2013 corpus is much larger (≈240 million *tweets*) than Tweets2011 (16 million *tweets*) used in TREC 2011 and TREC 2012. It was created by crawling Twitter's public sample stream over the period spanning from 1 February 2013 – 31 March 2013. The experiments were performed using both the query topics for the 2013 and 2014 editions of the TREC Microblog track [20, 21]. NIST provided relevance judgments TREC 2013 (60) and TREC 2014 (55) on a three-point scale: not relevant, relevant, and highly relevant.

5.2.2 External Collections: Twitter Verified Accounts. We crawled the timelines of Twitter's verified users (~205k accounts as of Aug 2016) collecting tweets from the period 1 February – 31 March 2013, which matches the period covered by the Tweets2013 TREC microblog dataset. Twitter's verified accounts belong to news organizations, mass media, and celebrities, so the posts have higher quality than a randomly sampled accounts. The cleaner vocabulary also allows the identification of interesting clusters more easily.

Topping the list of verified users (sorted by number of followers), there are a number of singers, actors, and other celebrities. Additionally, some accounts belong to companies that provide customer support through Twitter. These accounts provide customer support using private messages sent via Twitter Direct Messages (DMs). To be able to send DMs on Twitter, users have to follow each other. Thus, to help remove these two types of unwanted accounts, we extract two additional metrics for each account:

- the average number of tweets per day and
- the ratio between the number of replies and total posts.

To select high quality informative sources we remove accounts that meet the following criteria: $posts/day < 10$ and $\frac{replies}{posts} > $ $^1/_3$. Accounts that belong to news media outlets and other mass media organizations, typically produce a high volume of posts daily. Thus, we remove accounts that have a low daily average number of posts (e.g., @katyperry, @justinbieber, etc.). News accounts and broadcasters seldom reply to other users on Twitter, while accounts used by companies to provide customer support have a high ratio of replies (e.g., @XboxSupport, @AppleCare, etc.).

Each account's timeline is then classified in terms of written language by sampling their five most recent posts using ldig to remove non-English accounts. A total of 645 accounts were used, totaling approximately 800k *tweets.*

Tweets are tokenized using Twokenize[2], initially published alongside TweetMotif [25]. Preprocessing included removing URLs, email addresses, numbers, times, mentions, and emoticons. The tweets corpus was partitioned using mini-batch k-Means, with the number of clusters empirically set to $K = 200$ since the corpus covers a large period of 2 months.

5.3 Baselines and Experimental Systems

Relevance baselines. The first baseline is the query-likelihood retrieval model with Dirichlet prior smoothing [39] with $\mu = 2500$, which we will refer to as the **LM.Dir** model. The second strong baseline, **LTR**, is a learning to rank model combining multiple retrieval models (i.e., LM.Dir, BM25, IDF) and the features in Table 1.

Temporal baselines. There are three temporal ranking baselines: **Recency** [19], and **KDE(score)** and **KDE(rank)** [11], two different variants of a state-of-the-art temporal feedback method.

Experimental systems. The KDE_E method consists in performing temporal feedback on external collections as described in Section 3.1.1. The RM_E method uses the external collections, described in the previous section, to expand the initial query before searching the main corpus. The RMT_E experiment system uses time-based term expansion introduced in Section 3.2. Finally, the proposed $KDE+KDE_E+RMT_E$ experimental system uses both temporal vertical feedback and time-based term expansion. Whenever KDE is used, we opted for the KDE(rank) variant due to its better performance on previous publications. The learning to rank methods use *coordinate ascent* to optimize mean average precision (MAP).

6 RESULTS AND DISCUSSION

In this section we start by comparing the retrieval results of the different baselines, temporal methods, and the experimental systems, and then present a qualitative analysis of the temporal distribution of the results retrieved by different systems.

6.1 Retrieval Results

Time-based Relevance Models Using External Collections.
In this section we analyze the influence of time-based relevance models. The organization of the expansion corpus into topic-based verticals makes the query expansion process *temporally focused*. Verticals created by a partitioning algorithm using a topic-based similarity criteria exhibited different temporal profiles. The distribution of documents contained in each topic-based vertical is biased towards the time periods for when the vertical is most relevant. Following the temporal cluster hypothesis, the temporal relevance estimate extracted using the timestamps from the verticals selected was integrated into the retrieval process. In the pseudo-relevance feedback term selection stage it is used to generate *temporally focused* query expansion terms. In Table 2a and Table 2b we present a comparison of the results of MAP and P30 in the TREC 2013 and 2014 test topics. By estimating the relevance models using

---

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

[2]https://github.com/myleott/ark-twokenize-py

Table 2: TREC evaluation results.

<div style="display:flex">

(a) TREC 2013 dataset results.

Method	MAP	P30	Rprec
LM.Dir	0.2629	0.4622	0.3094
Recency	0.2663	0.4611	0.3115
KDE(score)	0.2583	0.4517	0.3004
KDE(rank)	$0.2736^{\dagger}$	$0.4878^{\dagger}$	$0.3178^{\dagger}$
LTR	0.2787	0.4617	0.3193
RM_E	0.2797	0.4528	0.3167
RMT_E	$0.2824^{\dagger}$	0.4700	0.3233
KDE_E	$0.2889^{\ddagger}$	$\mathbf{0.5061}^{\ddagger}$	$\mathbf{0.3322}^{\ddagger}$
$KDE+KDE_E+RMT_E$	$\mathbf{0.2900}^{\dagger}$	0.4850	0.3229

Symbols † and * stand for a $p < 0.05$ statistical significant improvement over KDE(score) and LTR respectively (‡ and $^{\ddagger}_{*}$ for $p < 0.01$).

(b) TREC 2014 dataset results.

Method	MAP	P30	Rprec
LM.Dir	0.4316	0.6315	0.4552
Recency	0.4323	0.6382	0.4576
KDE(score)	0.4205	0.6303	0.4476
KDE(rank)	0.4399	0.6406	0.4664
LTR	0.4469	0.6721	0.4625
RM_E	0.4705	0.6394	0.4890
RMT_E	$0.4738^{\ddagger}$	0.6442	$0.4927^{\dagger}$
KDE_E	$0.4643^{\ddagger}_{*}$	$0.6776^{\ddagger}$	$0.4869^{\dagger}_{*}$
$KDE+KDE_E+RMT_E$	$\mathbf{0.5183}^{\ddagger}_{*}$	$\mathbf{0.6970}^{\dagger}$	$\mathbf{0.5138}^{\ddagger}_{*}$

Symbols † and * stand for a $p < 0.05$ statistical significant improvement over KDE(score) and LTR respectively (‡ and $^{\ddagger}_{*}$ for $p < 0.01$).

</div>

the proposed time-sensitive term selection approach (RMT_E), the retrieval effectiveness always improved against the non-temporal method (RM_E). In fact, in TREC 2013 we observe a large effect of time-sensitive term selection on P30 when using the proposed vertical feedback architecture. Overall, we found that time-sensitive term selection is effective when used in standard pseudo-relevance feedback as well as in the proposed vertical feedback architecture.

Estimating Temporal Relevance Using External Collections. In this section we analyze the importance of temporal feedback from external collections. The major difference between KDE_E and $KDE+KDE_E+RMT_E$ is that the former uses the vertical feedback architecture for temporal feedback only, while the latter uses this architecture for query expansion via a time-aware pseudo-relevant vertical feedback method. In addition, it uses the estimate of temporal relevance obtained from temporal feedback on documents retrieved from the corpus using the expanded query. Like the LTR method, KDE_E is based only on the re-ranking of the documents retrieved by an initial retrieval method (i.e., LM.Dir). It is, therefore, very interesting that the KDE_E is not only very competitive against LTR and the KDE-based methods, but also with $KDE+KDE_E+RMT_E$. In the TREC 2013 queries, KDE_E even outperformed $KDE+KDE_E+RMT_E$ for both top-precision metrics, P30 and Rprec. $KDE+KDE_E+RMT_E$ outperformed the other methods on MAP, but the difference was not statistically significant against KDE_E. In the TREC 2014 queries the RMT_E-based methods outperform KDE_E on the recall-oriented metrics, MAP and Rprec. $KDE+KDE_E+RMT_E$ statistically significantly outperformed KDE_E in the recall-oriented metrics, MAP and Rprec, in part due to the use of the RMT_E method in $KDE+KDE_E+RMT_E$ to obtain the candidate set of documents for re-ranking.

Full Model Analysis. To conclude the retrieval results analysis, we examine the overall gains offered by temporal evidence from topic-based external collections. The results of the evaluation on the two TREC test datasets are summarized in Table 2a and Table 2b. We present the results for three retrieval effectiveness metrics: MAP, P30, and Rprec. We found that $KDE+KDE_E+RMT_E$ can outperform

non-temporal learning to rank models as well as state-of-the-art temporal ranking methods.

The $KDE+KDE_E+RMT_E$ method statistically significantly outperforms KDE(score) in both sets of queries. MAP improved 12.3% and 23.3% in the TREC 2013 and TREC 2014 topics respectively. Additionally, for the TREC 2014 topics the MAP result improved 17.8% over KDE(rank) and was statistically significant. Although in terms of P30, $KDE+KDE_E+RMT_E$ did not outperform KDE(rank) for the TREC 2013 queries, it outperformed KDE(score) albeit the result was not a statistically significant. In contrast, the improvements on P30 with $KDE+KDE_E+RMT_E$ on the TREC 2014 topics reached a statistically significant result of 10.6% over KDE(score) and 8.8% over KDE(rank), respectively.

$KDE+KDE_E+RMT_E$ outperforms the LTR baseline consistently across all metrics on both sets of queries. The improvements of $KDE+KDE_E+RMT_E$ in MAP and Rprec over LTR in the TREC 2014 topics were statistically significant, 16.0% and 10.0% for MAP and Rprec, respectively.

6.2 Temporal Distribution Analysis

This section aims to provide extra insights to understand the different performance of the retrieval methods in light of the effect on the temporal distribution of their top ranked documents. With this objective in mind, we look into a temporal representation of the *R-Precision* metric, Figure 2: we plot the ground-truth distribution of the R relevant documents of each query (empty bars) against the relevant documents retrieved at rank depth R (shaded bars). A perfect method retrieves only relevant documents, hence completely filling the empty bars. This visualization allows us to see if the methods are returning documents from the time periods that contain more relevant documents in the ground-truth. The plotted methods include the LM.Dir (no temporal evidence), KDE(rank) (temporal evidence from the corpus), KDE_E (temporal evidence from external collections), and $KDE+KDE_E+RMT_E$ (external, temporal feedback and time-based relevance model). Additionally, we present the EMD metric to quantify the difference between the temporal distribution of the retrieved documents and the true distribution. It is interesting to observe the direct relation between the EMD and Rprec results.

(a) MB124 – "celebrity DUI"

(b) MB133 – "cruise ship safety"

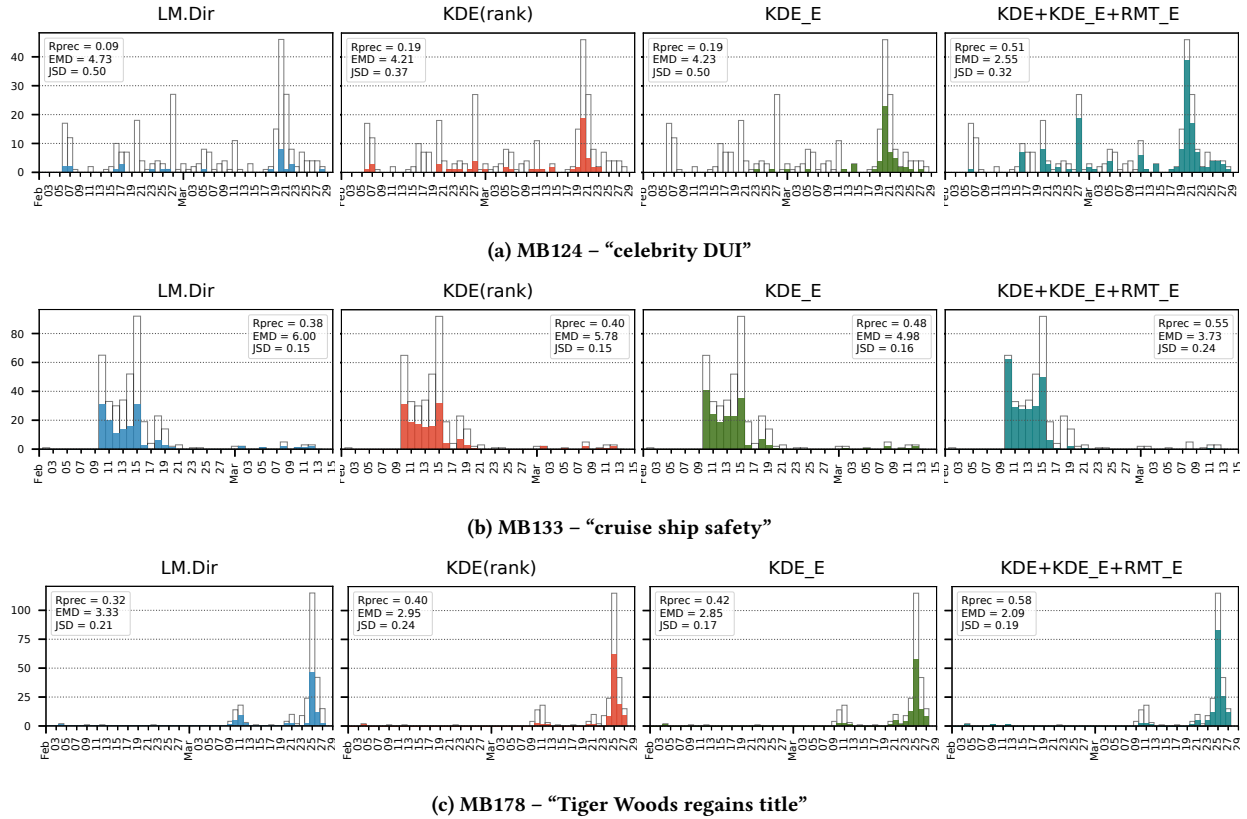

(c) MB178 – "Tiger Woods regains title"

Figure 2: Temporal profiles of queries and fit to the true distribution. The colored area of the bars represents the portion of relevant documents retrieved at a depth of R, where R is the number of relevant documents in the ground truth (i.e., Rprec).

In Figure 2, we plot some topics that improved the most. For all the queries shown, we observe that the temporal distribution of the top documents agrees with the temporal distribution of the documents in the ground truth.

For the top performing topic (see Figure 2a) we can see that KDE_E retrieves documents from the most relevant time period. However, with $KDE+KDE_E+RMT_E$ by using temporal query expansion, additional relevant time periods are found and retrieved. We can see that $KDE+KDE_E+RMT_E$ seems to retrieve more documents from the most relevant time period but it retrieves some documents from this second time period as well.

In the case of topic 133 "cruise ship safety", Figure 2b, it is clearly visible that $KDE+KDE_E+RMT_E$ is able to focus its retrieval towards documents published in February 10 and the following week. Inspecting the documents we found mentions to the Carnival Triumph cruise ship incident. This cruise ship set sail on February 7 and three days later (February 10) suffered an engine room fire.

The temporal distribution of the ground truth for topic 178 "Tiger Woods regains title", Figure 2c, indicates that most of the relevant documents are near the time of the query.

LM.Dir follows the temporal distribution of the ground truth. Nevertheless, the temporal distribution of the documents retrieved using KDE_E and $KDE+KDE_E+RMT_E$ shows that they can retrieve more documents from the most relevant days.

7 CONCLUSIONS

This paper presented the $KDE+KDE_E+RMT_E$ a time-aware and topic-aware pseudo-relevance feedback framework that mines textual and temporal signals from multiple information sources on Twitter. It explores the signals from verified accounts posts on Twitter, and temporal feedback to estimate the temporal relevance of search topics. The information streams from the verified accounts are automatically partitioned into verticals according to their topic.

Time-aware topical-based evidence mining. The results of the experiments confirmed our hypothesis that jointly modeling the topicality and temporality improves the estimation of relevance models, and yields improvements in Rprec along the timeline.

Efficient use of external collections. Building on recent advances, we show how to exploit the temporal heterogeneity of multiple external information verticals for time-aware ranking. These topic-based external verticals are exploited at two stages of the retrieval process: query expansion and time-aware ranking.

REFERENCES

[1] Robin Aly, Djoerd Hiemstra, and Thomas Demeester. 2013. Taily: Shard Selection Using the Tail of Score Distributions. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. ACM, New York, NY, USA, 673–682. https://doi.org/10.1145/2484028.2484033

[2] Jaime Arguello, Fernando Diaz, and Jamie Callan. 2011. Learning to Aggregate Vertical Results into Web Search Results. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*. ACM, New York, NY, USA, 201–210. https://doi.org/10.1145/2063576.2063611

[3] Michael Bendersky, Donald Metzler, and W. Bruce Croft. 2012. Effective Query Formulation with Multiple Information Sources. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*. ACM, New York, NY, USA, 443–452. https://doi.org/10.1145/2124295.2124349

[4] Jaeho Choi and W. Bruce Croft. 2012. Temporal Models for Microblogs. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 2491–2494. https://doi.org/10.1145/2396761.2398674

[5] Jaeho Choi, W. Bruce Croft, and Jin Young Kim. 2012. Quality Models for Microblog Retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 1834–1838. https://doi.org/10.1145/2396761.2398527

[6] Miguel Costa, Francisco Couto, and Mário Silva. 2014. Learning Temporal-Dependent Ranking Models. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 757–766. https://doi.org/10.1145/2600428.2609619

[7] Olga Craveiro, Joaquim Macedo, and Henrique Madeira. 2014. Query Expansion with Temporal Segmented Texts. In *Advances in Information Retrieval (ECIR '14)*. Springer, Cham, 612–617. https://doi.org/10.1007/978-3-319-06028-6_65

[8] Na Dai, Milad Shokouhi, and Brian D. Davison. 2011. Learning to Rank for Freshness and Relevance. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. ACM, New York, NY, USA, 95–104. https://doi.org/10.1145/2009916.2009933

[9] W. Dakka, L. Gravano, and P.G. Ipeirotis. 2012. Answering General Time-Sensitive Queries. *IEEE Transactions on Knowledge and Data Engineering* 24, 2 (Feb. 2012), 220–235. https://doi.org/10.1109/TKDE.2010.187

[10] Fernando Diaz and Donald Metzler. 2006. Improving the Estimation of Relevance Models Using Large External Corpora. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. ACM, New York, NY, USA, 154–161. https://doi.org/10.1145/1148170.1148200

[11] Miles Efron, Jimmy Lin, Jiyin He, and Arjen de Vries. 2014. Temporal Feedback for Tweet Search with Non-Parametric Density Estimation. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 33–42. https://doi.org/10.1145/2600428.2609575

[12] Weiwei Guo, Hao Li, Heng Ji, and Mona Diab. 2013. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 239–249.

[13] Rosie Jones and Fernando Diaz. 2007. Temporal Profiles of Queries. *ACM Trans. Inf. Syst.* 25, 3 (July 2007). https://doi.org/10.1145/1247715.1247720

[14] Nattiya Kanhabua and Kjetil Nørvåg. 2012. Learning to Rank Search Results for Time-Sensitive Queries. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 2463–2466. https://doi.org/10.1145/2396761.2398667

[15] Mostafa Keikha, Shima Gerani, and Fabio Crestani. 2011. Time-Based Relevance Models. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. ACM, New York, NY, USA, 1087–1088. https://doi.org/10.1145/2009916.2010062

[16] Anagha Kulkarni, Almer S. Tigelaar, Djoerd Hiemstra, and Jamie Callan. 2012. Shard Ranking and Cutoff Estimation for Topically Partitioned Collections. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 555–564. https://doi.org/10.1145/2396761.2396833

[17] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What Is Twitter, a Social Network or a News Media?. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, New York, NY, USA, 591–600. https://doi.org/10.1145/1772690.1772751

[18] Victor Lavrenko and W. Bruce Croft. 2001. Relevance Based Language Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*. ACM, New York, NY, USA, 120–127. https://doi.org/10.1145/383952.383972

[19] Xiaoyan Li and W. Bruce Croft. 2003. Time-Based Language Models. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM '03)*. ACM, New York, NY, USA, 469–475. https://doi.org/10.1145/956863.956951

[20] Jimmy Lin and Miles Efron. 2013. Overview of the TREC-2013 Microblog Track. In *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*, Ellen M. Voorhees (Ed.), Vol. Special Publication 500-302. National Institute of Standards and Technology (NIST).

[21] Jimmy Lin, Miles Efron, Yulu Wang, and Garrick Sherman. 2014. Overview of the TREC-2014 Microblog Track. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. Special Publication 500-308. National Institute of Standards and Technology (NIST).

[22] Flávio Martins, João Magalhães, and Jamie Callan. 2016. Barbara Made the News: Mining the Behavior of Crowds for Time-Aware Learning to Rank. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16)*. ACM, San Francisco, CA, USA.

[23] Flávio Martins, João Magalhães, and Jamie Callan. 2018. A Vertical PRF Architecture for Microblog Search. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '18)*. ACM, New York, NY, USA.

[24] Donald Metzler, Congxing Cai, and Eduard Hovy. 2012. Structured Event Retrieval over Microblog Archives. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 646–655.

[25] Brendan O'Connor, Michel Krieger, and David Ahn. 2010. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In *Fourth International AAAI Conference on Weblogs and Social Media*. 384–385.

[26] Maria-Hendrike Peetz, Edgar Meij, and Maarten de Rijke. 2013. Using Temporal Bursts for Query Modeling. *Information Retrieval* (July 2013), 1–35. https://doi.org/10.1007/s10791-013-9227-2

[27] Jinfeng Rao and Jimmy Lin. 2016. Temporal Query Expansion Using a Continuous Hidden Markov Model. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (ICTIR '16)*. ACM, New York, NY, USA, 295–298. https://doi.org/10.1145/2970398.2970424

[28] Tetsuya Sakai. 2014. Statistical Reform in Information Retrieval? *SIGIR Forum* 48, 1 (June 2014), 3–12. https://doi.org/10.1145/2641383.2641385

[29] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. 2009. TwitterStand: News in Tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '09)*. ACM, New York, NY, USA, 42–51. https://doi.org/10.1145/1653771.1653781

[30] Milad Shokouhi. 2007. Central-Rank-Based Collection Selection in Uncooperative Distributed Information Retrieval. In *Advances in Information Retrieval (ECIR '07)*. Springer, Berlin, Heidelberg, 160–172. https://doi.org/10.1007/978-3-540-71496-5_17

[31] Milad Shokouhi and Luo Si. 2011. Federated Search. *Found. Trends Inf. Retr.* 5, 1 (Jan. 2011), 1–102. https://doi.org/10.1561/1500000010

[32] Luo Si and Jamie Callan. 2003. Relevant Document Distribution Estimation Method for Resource Selection. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR '03)*. ACM, New York, NY, USA, 298–305. https://doi.org/10.1145/860435.860490

[33] Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. 2011. Linking Online News and Social Media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*. ACM, New York, NY, USA, 565–574. https://doi.org/10.1145/1935826.1935906

[34] Yulu Wang and Jimmy Lin. 2017. Partitioning and Segment Organization Strategies for Real-Time Selective Search on Document Streams. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*. ACM, New York, NY, USA, 221–230. https://doi.org/10.1145/3018661.3018727

[35] Wouter Weerkamp, Krisztian Balog, and Maarten de Rijke. 2012. Exploiting External Collections for Query Expansion. *ACM Trans. Web* 6, 4 (Nov. 2012), 18:1–18:29. https://doi.org/10.1145/2382616.2382621

[36] Stewart Whiting, Iraklis A. Klampanos, and Joemon M. Jose. 2012. Temporal Pseudo-Relevance Feedback in Microblog Retrieval. In *Advances in Information Retrieval (ECIR '12)*. Springer, Berlin, Heidelberg, 522–526. https://doi.org/10.1007/978-3-642-28997-2_55

[37] Tan Xu, Douglas W. Oard, and Paul McNamee. 2014. HLTCOE at TREC 2014: Microblog and Clinical Decision Support. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. Special Publication 500-308. National Institute of Standards and Technology (NIST).

[38] Chengxiang Zhai and John Lafferty. 2001. Model-Based Feedback in the Language Modeling Approach to Information Retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM '01)*. ACM, New York, NY, USA, 403–410. https://doi.org/10.1145/502585.502654

[39] Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Trans. Inf. Syst.* 22, 2 (April 2004), 179–214. https://doi.org/10.1145/984321.984322