Leveraging Temporal Query-Term Dependency for Time-Aware Information Access

Bilel Moulahi*†, Lynda Tamine* and Sadok Ben Yahia†‡
*IRIT, University of Toulouse, France
†Faculty of Science of Tunis, University of Tunis El Manar, Tunisia
‡Institute Mines-TELECOM, TELECOM SudParis, UMR CNRS Samovar, France
*{moulahi,lechani}@irit.fr, †sadok.benyahia@fst.rnu.tn

***Abstract*—Incorporating the temporal property of queries into time-aware information access methods has been shown to have a significant positive effect on a large number of search tasks, such as over microblogs and news archive. Recent work on time-aware search mostly rely on time-based relevance models that are built upon the language model framework. However, in this model, query terms are often assumed to be generated independently from each other. In this paper, we observe through a time series analysis that, query terms are temporally dependent and are frequently occurring within similar time periods when they deal with the same topics. In contrast to existing work, we propose a method that naturally extends the effective temporal language model and exploits this dependency at the term granularity level. Moreover, we reframe the task as a rank aggregation problem that fully exploits the temporal features of query terms. Experiments using the large-scale TREC Temporal Summarization 2013 and 2014 standard datasets empirically show that our method leads to significant performance improvements, when compared to state-of-the-art temporal ranking models.**

I. Introduction

Temporal information retrieval (IR) [1] is a new emerging research field that aims to retrieve temporal relevant documents. In the last decades, several effective IR models have exploited temporal relevance criteria to enhance the retrieval effectiveness [2], [3], [4]. These approaches went beyond simple topical relevance matching and incorporated temporal properties of documents and query words into the ranking models. Time is thus represented by a variety of relevance features, such as query recency and document freshness, to satisfy temporal information needs [5], [6], [7], [8]. Though these methods are shown to have significant impact on retrieval effectiveness, they do not fully exploit the temporal information hidden behind the query terms and documents. Indeed, most of time-aware information access methods rely on the language model framework, in which documents and especially queries are represented as bag of words, and document relevance is often generated for the whole query. Accordingly, the query terms are often assumed to be generated independently from each other. However, the latter may be temporally dependent and this correlation may be used as a temporal signal to adjust the final document ranking. Looking at the temporal characteristic of each term aside opens the door towards identifying the time periods of interest for the topic. For instance, people tend to talk about the "*uefa champions league*" mainly during or slightly after time periods when the tournament was held (qualifying, competition proper and group stage). Thus, documents created beyond these time periods are less likely to discuss the UEFA competition, even if they contain one of these terms. Hence, they are less likely to be relevant. Figure 1 illustrates this observation; it shows the

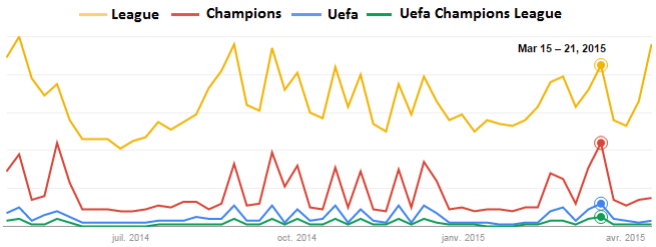


Fig. 1: Interest over time on the query terms "uefa", "champions", "league" as well as the whole query "uefa champions league", from Google trend (accessed on April 2015). The temporal distribution is given throughout this year.

interest over time of the query "*uefa champions league*", as given by Google trend. This graph clearly illustrates that the event jointly triggers an increase of interest for the three query terms *"uefa", "champions"* and *"league"* in very particular time windows (e.g., from Mars 15 to Mars 21, 2015, as highlighted in the Figure). This leads to the assumption that documents, that are topically relevant for all the query terms and published in these *bursty* time periods, are more likely to be relevant in response to the whole query.

With this in mind, we address the document ranking problem from this temporal dependency perspective. We propose a time-aware ranking approach aggregating the topical relevance matching criterion with a temporal relevance factor, based on a query term cross-correlation analysis. Our model is based on an effective existing temporal language modelling framework that provides a principled means to combine these criteria [6]. In particular, we apply this model at the query-term granularity level, so that each term is considered as a separate query. More specifically, each query term is assumed to return a temporally-dependent document ranking list. Then, we reframe the task of computing the global query-document matching as a rank aggregation problem [9]. To model the temporal query-term proximity, we extend an existing rank fusion method [10], in order to boost documents that are published in the same time periods as a large number of relevant documents. This fits with our intuition that relevant documents are those published in *bursty* time periods and highly ranked in many of the query-terms ranking lists. To the best of our knowledge, there is

no prior attempts to tackle the temporal term dependency in such a way. We evaluate the proposed time-aware ranking approach in the context of TREC Temporal Summarization task. Our experiments focus on the value that temporal term cross-correlation can bring to time-aware document retrieval. We provide a detailed analysis of the performance of our approach compared to existing temporal ranking methods.

The remainder of the paper is organized as follows. In Section II, we discuss the prior related work. In Section III, we introduce some motivations and present the research questions. Then, we present our approach for time-aware information access in Section IV. In Section V we set up our experimental environment and evaluate the proposed time-aware document ranking model in comparison with state-of-the-art both topical and time-aware document ranking models. In Section VI, we conclude the paper and report some future work.

II. Related Work

Two main lines of research are close to our work: time-aware IR and rank aggregation.

A. Time-Aware Information Retrieval

Temporal IR is a new emerging research field that aims to enhance the retrieval effectiveness by embedding temporal characteristics of queries and documents in the core of the ranking process. Previous work show that a large amount of web documents become time-dependent [11], [12]. Metzler et al. [3] have argued that about 7% of queries have implicit temporal intent, while other studies show that only 1.5% of queries are explicitly provided with temporal intent [13]. To exploit the temporal information contained in these queries and documents, most state-of-the-art studies were based on language modelling frameworks and linear combination mechanisms to aggregate temporal and content matching criteria [4]. Li and Croft [8] examined the temporal distribution of documents to classify queries and exploited this information within a temporal language model. Two types of time-based queries are identified, the first favours the most recent documents and the other is shown to have relevant documents within a specific period in the past. For the first type of queries, most recent documents obtain the higher probabilities of relevance. Results on TREC ad-hoc queries show that, for a specific set of recency queries, time-based query likelihood language models outperform query likelihood language models and linear combination reranking methods. Efron and Golovchinsky [6] proposed a Bayesian estimation based methods for incorporating time into language modelling framework. They integrate a query-specific estimation information into an exponential distribution to introduce a penalty for the document age. They showed that their method significantly outperforms state-of-the-art approaches for recency queries, on Twitter data. In a close work, Massoudi et al. [7] proposed a query expansion model for microblogs, which selects terms temporally closer to the query submission time. Their model is supposed to work well for finding documents related to events currently happening but, not as well for past events. Following the same direction, Dakka et al. [4] proposed a general language model that incorporates time into the ranking model in a principled manner. For a given time-sensitive query over a news archive, the approach automatically identifies significant time intervals for the query and uses them to adjust the document relevance scores by boosting the scores of documents published within the important intervals [4]. Other work attempt to explore the relationship existing between time and relevance in query expansion settings. For instance, Metzler et al. [14] proposed a temporal query expansion method for microblogs based on the temporal co-occurrence of terms in a timespan. They first performed pseudo-relevant timespan retrieval for an event query (e.g., earthquake) and used those timespans for query expansion. Although their goal was retrieving a ranked list of historical event summaries, the temporal query expansion method showed that selecting relevant timespans is crucial for query expansion for microblogs.

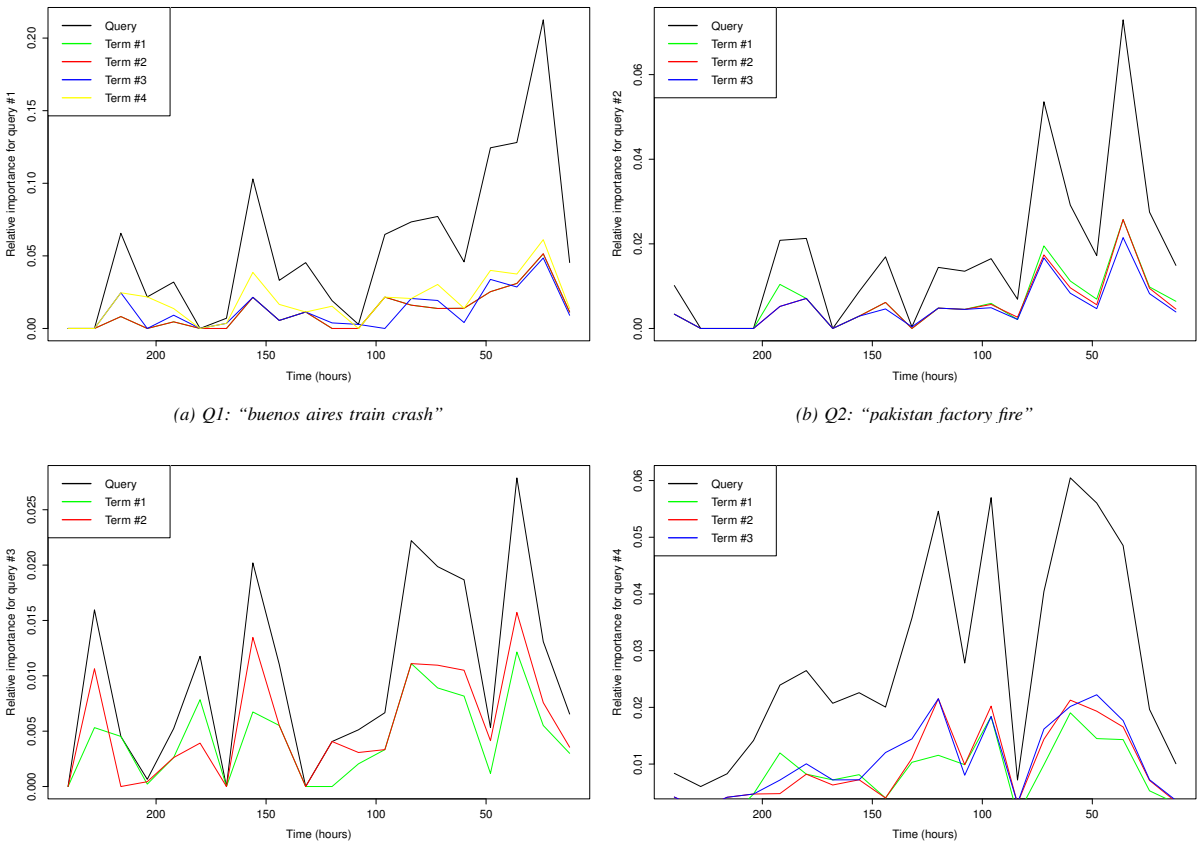
B. Rank Aggregation

The rank aggregation problem consists in combining many ranking results, obtained from different individual rank functions, in order to obtain a better ranking list [9]. These ranking fusion methods can be classified based on whether they rely on the scores or the ranks of documents in the different lists [15]. These methods can also be categorized into supervised and unsupervised methods, depending on the use of a training model. Researches on this realm have attracted much attention in the last decades and have been successfully applied in many research areas such as meta-search [9], [16], word association finding [17] and similarity search [18]. The most widely used rank aggregation methods include but not restricted to the *Comb** family, outranking approaches [19], Reciprocal Rank Fusion methods (RRF) [10], fusion methods such as Condorcet [20] and Borda count [9]. Most of these methods assume that only documents that are highly ranked in many of the rankings are more likely to be relevant. Compared to some established standard rank aggregation approaches, the RRF method has been shown to be very effective [10] within TREC collections. Cormack et al. [10] show that RRF consistently equalled or bettered some established state-of-the-art fusion methods such as Condorcet Fuse and CombMNZ [20]. The authors claim that RRF is better able to harness diversity within individual rankings, unlike Condorcet, in which a simple majority of weak preferences may substantially overrule stronger ones [10]. Despite the extensive body of work done in this scope, none of these methods attempt to model dependencies between the ranking lists to be fused. The ranks and scores are treated independently by means of sums or pruning techniques to discard the bad rank functions altogether. In this paper, we extend the RRF method and apply it at the query level, to close the gap between the temporal properties of queries and the dependency that might exist among the query-terms.

III. Motivations and Research Questions

We plot in Figure 2, the time-series analysis [21], [22] curves of four test queries (Q1-Q4) as well as their within terms, in the relevant documents (qrels) of the TREC 2013 Temporal Summarization track[1]. The $x-$axis represents time in hours (document ages from query time to document timestamps), and the $y-$axis indicates the normalized weight of the query (and terms) over the documents. The time series are constructed from the set of relevant documents for the four queries. We examined the collection at hourly intervals,

[1] https://www.trec-ts.org



(a) Q1: "buenos aires train crash"

(b) Q2: "pakistan factory fire"

(c) Q3: "colorado shooting"

(d) Q4: "sikh temple shooting"

Fig. 2: Time-series analysis of queries Q1-Q4 and their within terms in the true relevant documents (qrels) in the TREC 2013 Temporal Summarization track.

which form the figure's x-axis. Figure 2 clearly shows that the temporal distribution of most of the relevant documents are clustered into specific time periods, for all the studied queries. Interestingly, this distribution is substantially similar across the within query terms. Our basic intuition, is that this cross temporal dependency provides us with useful information about the relevancy of documents. These documents are those occurring in specific time periods and that are jointly relevant for all the query-terms. For instance, in Figure 2b, the terms of the query "pakistan factory fire" have the same temporal distribution within many time periods (e.g., the first 50 hours). An interesting point that may be raised is when a term occurs in many queries. This holds for the term "*shooting*" (term $\#2$ in $Q3$ and term $\#3$ in $Q4$). Unfortunately, in this case, the query submission times of both queries are totally different and do not overlap, but if it was the case the term would particularly exhibit different behaviours in some time frames compared to the other terms. We believe that taking into account the temporal dependency with the other within query-terms would compensate the local inaccuracies of the single term, given that single terms alone often fail to provide a clear idea about related relevant documents. This is counterpart to term-proximity measures, but with an emphasis on the temporal perspective.

In the following, we attempt to exploit this insight to answer two research questions:

1) **RQ1**. In what extent query-terms are temporally dependent within relevant documents?
2) **RQ2**. How can we exploit query-terms temporal correlation to enhance time-aware document retrieval?

IV. Using Temporal Term-dependency for Time-Aware Information Access

Our approach for time-aware information access is built upon an effective time-based model that integrates time factors into a language modeling framework. Thus, we first introduce as a stepping stone the basic query likelihood model, followed by a presentation of the temporal language model. Then, we describe our proposed approach and show how it exploits term-query correlation into these frameworks.

A. Problem

We first start by formalizing the problem and introducing some basic definitions and notations. Suppose we have a query $q = w_1, w_2, \ldots, w_n$, where w_i is a query term, and a document $d_j^t \in \mathcal{D}$, where t is the time publication of d_j. t is a discrete timestamp that indicates the number of time units (e.g., hours)

since a reference time (e.g., the UNIX epoch). The problem consists in bringing up "relevant" documents published around times of interest to query q. The main contribution of this paper is the exploitation of the temporal signals contained in the query-terms level. As the relevance is computed on the query-terms level, the first challenge concerns the computation of the single query-terms relevancy with respect to both topical and temporal matching criteria.

B. The Model

Our model includes two main steps, illustrated in Algorithm 1. The first step consists in computing the single query-terms relevancy with respect to a topical matching criterion $P(w_i|d_j)$ and a temporal relevance model $P(t|w_i)$. This leads to a number of ranked lists associated with each query term. Then, we identify the time-span of the top K highly ranked documents of each result list. We define a set of important periods for all of these documents. We estimate this time period as the average of the top K document timestamps returned wrt the query terms. In the second step, according to our intuition detailed above, we merge the ranked lists into one ranking result. The goal of this step is to favour documents that are published in the same time periods as a large number of relevant documents that are returned in response to all of the query-terms. Table I describes the notations used within Algorithm 1.

Notation	Description
n	Number of terms in a query
q	A query which contains n terms $(w_1, w_2, \ldots, w_n)$
d^t	A document published at time t
r_{w_i}	A document ranking returned wrt a term w_i, $r_{w_i} \in R$
R	The set of ranking lists wrt the query terms

TABLE I: A summary of notations used within Algorithm 1.

Algorithm 1 The Model

Data: $q = w_1, w_2, \ldots, w_n$, n, $d_j^t \in \mathcal{D}$.
Result: Ranks of documents
Step 1: compute the single query-terms relevancy
1. **For** $i = 1$ to n {*Each query-term is considered as a query*} **do**
2. Compute the topical matching criterion $P(w_i|d_j^t)$ ($j \in 1 \ldots |\mathcal{D}|$)
3. Compute the temporal relevance score ($P(t|w_i)$)
4. Compute the global score of each document (Combine topical and temporal scores)
5. r_{w_i} := Top K documents returned wrt w_i
6. **End for**

Step 2: Rank aggregation
7. Identify the average time-span of r_{w_i} $(i : 1 \ldots n)$
8. Fuse the ranked lists of each query-term
9. **Return** Document ranking of query q

1) Generating the Query-Terms Rankings: Our approach relies on the time-based model proposed in [4], based in turn on a temporal probabilistic model from [8]. We start by applying the model on the query-term granularity level, i.e., each term is individually viewed as a query. The proposed model ranks documents in decreasing order of their probability of relevance based on their temporal ($P(t|w_i)$) and topical ($P(q|d)$) relevance:

$$P(d^t|w_i) = P(d, t|w_i) \propto P(d|w_i)P(t|w_i) \quad (1)$$
$$\propto P(q|w_i)P(d)P(t|w_i) \propto P(w_i|d)P(t|w_i) \quad (2)$$

Where $P(w_i|d)$ denotes the query-term likelihood on document d, $P(d)$ stands for the prior probability that d is relevant to any query-term. Since $P(d)$ is uniform, it is discarded from the formula. Given that w_i consists of a lexical term, $P(w_i|d)$ can be estimated using a standard text-based query likelihood method. To mitigate the problem of the zero-probability problem, if the query term has zero probability of being generated from the document, we use the Dirichlet smoothing, yielding:

$$P(w_i|d) = \frac{tf(w_i, d) + \mu.\frac{tf(w_i,d)}{|D|}}{|d|+\mu} \quad (3)$$

where $tf(w_i, d)$ stands for the frequency of w_i along d.

The second factor $P(t|w_i)$ conveys the relative importance of the time point t for the query-term w_i. This temporal relevance can be estimated using different methods, e.g., the maximum likelihood model, which is defined as the normalized sum of the relevance scores of documents published at time t for query-term w_i:

$$P(t|w_i) = \frac{tf(w_i, D^t)}{|D^t|} \quad (4)$$

where D^t is the set of documents published at time t. This weighting function assumes the temporal independency of the query terms.

2) Time-Aware Rank Aggregation of Query-terms Rankings: The query-terms generated rankings obtained in step 1, give rise to different lists $r_w \in R$ wrt both topical and temporal criteria. To merge these lists, we extend an existing RRF rank fusion method [10] by injecting a temporal proximity distance that exploits the temporal term dependency. To characterize this temporal proximity, we apply the normalized variant of the so-called Gaussian kernel function. The documents scores given by our Temporal Term Dependent Model (TTD-M) are computed as follows:

$$TTDM(d^t \in D) = \sum_{r \in R} \frac{1}{\epsilon + r(d_t)} * kernel(t, t_{avg}) \quad (5)$$

where $r_w(d^t)$ is the position of document d in the rank list r_w and t_{avg} is the average time of the top highly ranked documents in R. We assumed that t_{avg} is the most important time period for a given query. This rewards documents, returned by all (or most of) the query terms, that are published closer to time frame of the K highly ranked documents. That is, if two documents are close enough in terms of importance and time, for all the query terms, they should be highly ranked.

The Gaussian density function is computed as follows:

$$kernel(t_1, t_2) = \frac{1}{\sqrt{2\pi}\sigma} * exp[\frac{-(t_1 - t_2)^2}{2\sigma^2}] \quad (6)$$

where σ refers to the variance of the density kernel.

V. Experimental Evaluation

In this section, we detail our experimental evaluation. First, we describe the experimental setup; we present the used dataset, baselines and evaluation metrics. Then, we report the results of (i) a cross-temporal analysis over the query-terms as well as (ii) a comparative retrieval effectiveness with standard temporal as well as atemporal document retrieval methods.

A. Experimental Setup

1) Datasets and Evaluation Metrics: We used the TREC KBA 2013 and 2014 Stream Corpus[2] officially exploited by TREC Temporal Summarization (TS) track[3]. The TREC KBA Stream Corpus consists in a set of over 2 billions timestamped documents from a variety of news and social media sources covering the time period October 2011 to January 2013. Each document contains a set of sentences, each with a unique identifier. A set of 10 topics is included in this dataset, where each query refers to a real world event and is given a starting and ending time. The TREC TS Track aims to return relevant documents about these events. It consists of two tasks: (i) *Sequential Update Summarization* where participants should return relevant and novel sentences (updates) for a developing event and (ii) *Value Tracking* that intend to estimate values for a particular attribute for an event (e.g., the number of fatalities or financial impact). Both sub-tasks have clear temporal character since the updates have to be timely and relevant. In this work, we consider the *Sequential Update Summarization* task. Given that we are interested in the retrieval of documents, we perform all the tests on the document level rather than the sentence level, i.e., a document is considered as relevant if it contains some relevant sentences. We evaluate our approach as well as the baselines using the measures of precision, recall and F-measure computed as follows:

$$Precision = \frac{|\text{Relevant retrieved documents}|}{|\text{Retrieved documents}|}, \; Recall = \frac{|\text{Relevant retrieved documents}|}{|\text{Relevant documents}|}$$

$$F\text{-}measure = 2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}}$$

We conducted the correlation analysis on the TREC 2013 TS queries and documents, then we evaluated the effectiveness of our document ranking model as well as the baselines using the data of the TREC 2014 TS track. We indexed the document collections using Lucene[4]. The index is created to simulate a realistic real-time search setting, where no future information is available when a query is issued. As the corpus of these timestamped documents is considered as a stream, we build an index for each hour. In Figure 3, we present an example of query ($Q1$) which corresponds to the event "Hurricane sandy".

2) Baselines: We compare our document ranking model to an atemporal document ranking model: (i) LM as the unigram language model with Dirichlet smoothing [23] (as presented in Eq. 3) and two temporal ranking models: (ii) TLM as the temporal language model by Dakka et al. [4] (according to Eq. 1), and (iii) the Recency Prior (RP) by Li and Croft [8].

[2]http://trec-kba.org/kba-stream-corpus-2013.shtml
[3]http://www.trec-ts.org/
[4]https://lucene.apache.org/

```
<event>
    <id>1</id>
    <title>2012 Buenos Aires Rail Disaster</title>
    <description>http://en.wikipedia.org/wiki/2012_Buenos_
    Aires_rail_disaster</description>
    <start>1329910380</start>
    <end>1330774380</end>
    <query>buenos aires train crash</query>
    <type>accident</type>
</event>
```

Fig. 3: Query Q1 "buenos aires train crash" of the TREC 2013 TS track.

The RP defines a prior distribution over documents to favour recent documents:

$$P(d) = \lambda e^{-\lambda t_d} \tag{7}$$

where λ is the rate of an exponential distribution and t_d is the age of document d.

Table II presents some statistics about the queries and their corresponding relevant documents.

Query *Id*	Query	#Relevant Documents
1	Q1: Buenos Aires Crash Train	789
2	Q2: Factory Fire Pakistan	585
3	Q3: Colorado Shooting	243
4	Q4: Shooting Sikh Temple	613
5	Q5: Hurricane Isaac	36
6	Q6: Hurricane Sandy	518
7	Q7: Derecho midwest	2
8	Q8: Bopha Typhoon	210
9	Q9: Earthquake Guatemala	294
10	Q10: Tel Aviv Bombing Bus	284

TABLE II: TREC TS 2013 Queries and their corresponding number of relevant documents.

B. Cross Temporal Query-Term Analysis

According to our first research question (RQ1), we first investigated to what extent the query-terms are dependent in order to validate our intuition about the temporal query-terms correlation. Figure 4 depicts the similarity matrix representing the cross-temporal correlation between the time series of the TREC 2013 TS query terms. The temporal cross-correlation measures the similarity between two time series [21] based on the relative importance of each term (cf., Eq. 4). Both columns and rows of the matrix refer to the terms belonging to the queries (Q1-Q10), here a total of 26 terms. Query terms are ordered by query *Id* in both the x-axis and y-axis. Accordingly, the blocks in the diagonal correspond to the similarity sub-matrix that involves the terms belonging to the same query. The *Id* of each query represented within each block is highlighted in bold. A matrix cell highlights, using colour intensity, the degree of correlation between the corresponding terms. While the blue colour means a high similarity, the red indicates an opposite low similarity. More brighter the colour, less intense is the correlation measure. We can clearly see from Figure 4 (based on the colour intensity as indicated above) that the temporal series of terms belonging to the same query are more correlated than the temporal series of terms not belonging to the same query (out of the blocks). This observation corroborates our assumption about the temporal

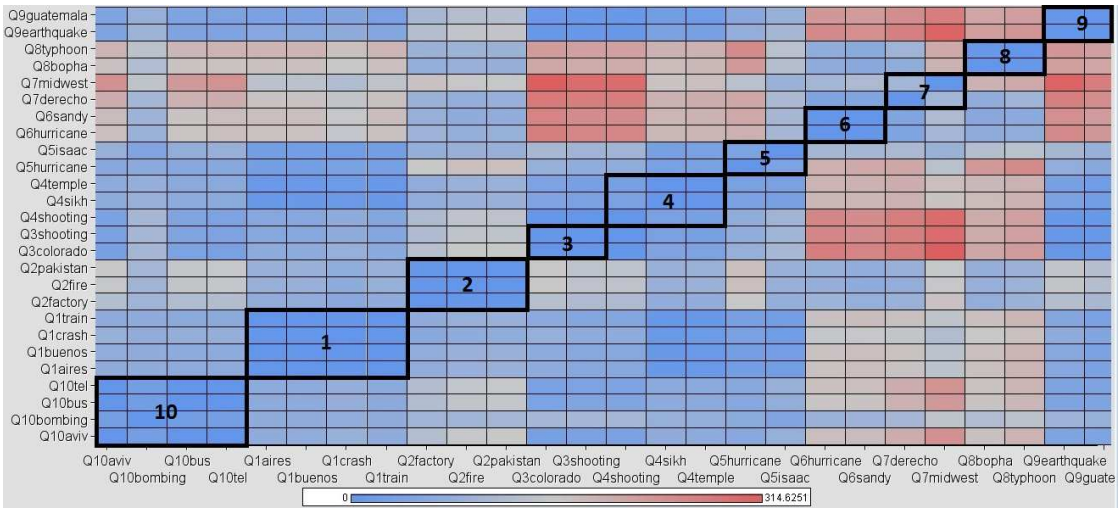


Fig. 4: Cross-temporal analysis of the TREC 2013 TS query-terms.

dependency of within-terms in comparison with other terms. Interestingly, we can notice that terms "midwest" ($term\#1$) and "derecho" ($term\#2$) of query Q7 are not temporally correlated. This observation could be explained either by the fact that a number of relevant documents wrt other queries occurs after Q7 submission time or by the low number of relevant documents for query Q7, as shown in Table II.

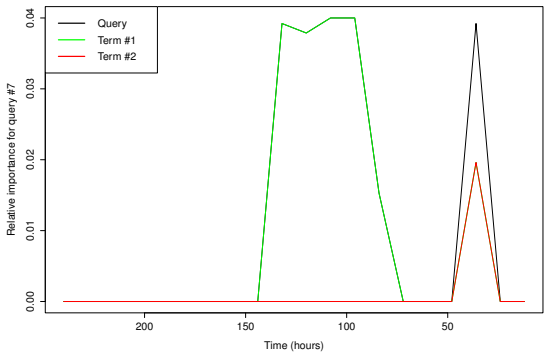


Fig. 5: Distribution over time (measured by hours) of query Q7 ("midwest derecho") and its within terms in the true relevant documents the TREC 2013 TS track. The $x-$axis represents time in hours and the $y-$axis indicates the normalized weight of the query (and terms) over the documents.

To get a deeper understanding of this observation, we plotted in Figure 5 the temporal distribution of the terms belonging to query $Q7$ "midwest derecho" and its within terms in the true relevant documents provided in the TREC 2013 TS track. Figure 5 shows that the term "midwest" ($term\#1$) occurs in a bulk of relevant documents between the first 60 hours and 150 after the event Q7 occurrence. Indeed, the term "midwest" is not a keyword for this particular event, so it can also be frequent in relevant documents in response to other queries.

C. Effectiveness Evaluation

Our objective here is to answer the second research question (RQ2). To achieve this goal, first, we set up the parameters of our model as well as the baseline parameters. The smoothing parameter μ of the Language Model is set with a value of 2000, the rate parameter λ of the recency prior model is set to 0.01 (similar to Efron and Golovchinsky [6]). The parameter ϵ of the RRF fusion method and the Gaussian Kernel parameter σ are empirically set to 30 and 170, respectively. In each experiment, we first use the Dirichlet Smoothing model to retrieve 100 documents from each hour (index), in response to each query. Then, we use the other models to re-rank them.

Second, using the TREC TS 2014 dataset, we compared the effectiveness of our Temporal Term Dependency based Model (TTD-M) to the baselines listed above, namely, the atemporal retrieval model (LM) and two temporal retrieval models TLM and Recency Prior (RP). Table III reports the obtained results by means of Precision, Recall and F-Measure, the improvements and their significance based on a t-test.

	Precision	Recall	F-Measure	% Change
LM	0.0830	0.2019	0.1177	+32.47% §
TLM	0.1307	0.1772	0.1504	+13.71% §
RP	0.0866	0.2019	0.1212	+30.46%
TTD-M	**0.1692**	**0.1797**	**0.1743**	-

TABLE III: Comparative evaluation of retrieval effectiveness of our Temporal Term Dependency based Model (TTD-M). The last column (% change) indicates the improvements in terms of F-Measure. The symbols § denotes the student test significance: "§": $t < 0.05$.

We can observe from Table III that our approach significantly outperforms the language modelling (LM) framework with an improvement of about +32.74%. This result is in some extent not surprising given that the LM method is based only on a topical criterion, and ignores the temporal one. As previously mentioned, the data collection is strongly dependent on time, and thus using the topical matching alone could misestimate the relevance score of some documents.

Performance improvement with respect to the TLM document ranking model is less important with an F-measure value of 0.15. As shown in Eq.1, a temporal score is computed for each document. This score boosts the documents published in times of interest to the query.

In addition, when we compare the TTD-M method to the RP ranking model, we can see that TTD-M performs better, but the difference is not significant. This could be explained by the fact that relevant documents of the queries are quite uniformly distributed over different time periods. That is, there are queries for which the recency is not a major requirement of a user's information need. Thus, favouring only most recent documents could be detrimental for the retrieval performance. In order to better understand the results, we present in Table IV a query-level analysis of the performance of the TTD-M model in comparison to the RP model for each query of the TS 2014 topics.

Id	Query terms	F-Measure		
		TTD-M	RP	%change
11	costa concordia	0,2055	0,0904	**+55,98%**
12	european cold wave	0,0763	0,0347	**+54,49%**
13	queensland floods	0,2262	0,0787	**+65,21%**
14	boston marathon bombing	0,0802	0,1171	-45,99%
15	egyptian riots	0,1525	0,1028	**+32,56%**
16	quran burning protests	0,3646	0,2352	**+35,47%**
17	in amenas hostage crisis	0,1252	0,2361	-88,59%
18	russian protests	0,2107	0,0971	**+53,89%**
19	romanian protests	0,3470	0,0794	**+77,10%**
20	egyptian protests	0,0831	0,0727	**+12,48%**
21	russia meteor	0,0707	0,143	-100%
22	bulgarian protests	0,1967	0,0606	**+69,15%**
23	shahbag protests	0,0281	0,0489	-73,92%
24	nor'easter	0	0	0
25	Southern California shooting	0,0057	0,0510	-100%

TABLE IV: Query-level analysis of retrieval effectiveness of our Temporal Term Dependency based Model (TTD-M) vs. the RP model. The last column (% change) indicates the improvements in terms of F-Measure.

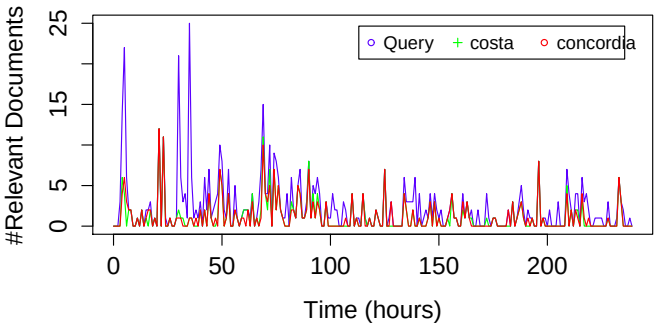


Fig. 6: Time-series of query Q11 ("costa concordia") and its within terms in the true relevant documents of the TREC 2014 TS track.

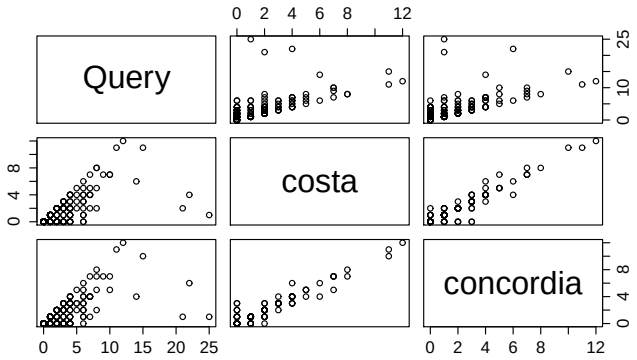


Fig. 7: Query-terms correlation of query Q11 and its within terms ("costa" and "concordia") over time (in hours).

Table IV shows that TTD-M outperforms RP for 60% of the queries (9/15). An analysis of queries with positive improvement shows that most of their corresponding relevant documents are quite uniformly distributed over time with some spikes in specific time periods. For instance, Figure 6 shows the temporal distribution of the true relevant documents for query Q11 of Table IV. This Figure clearly illustrates our observation. Unlike the RP model, the TTD-M method exploits the temporal factor as for the TLM, and extends the latter to favour documents that exhibit a similar relative importance shape over time for all the query-terms as shown in Figure 6. We conjecture that, by using the RRF method in conjunction with the Gaussian kernel density function, we are able to improve the ranking of popular documents across the different ranking lists wrt each query-term. All documents occurring in different lists, in the same time window, are highly ranked in the final ranking process. To confirm this query-terms cross-temporal correlation property which is our initial assumption in this paper, we plot in Figure 7 the correlation matrix of query Q11 and its within terms. Figure 7 shows that both terms "costa" and "concordia" are temporally correlated with a correlation value of 0.96. The correlation of the terms with the whole query is also quite significant with a value of 0.6725 and 0.6702 for "costa" and "concordia", respectively.

On the other side, TTD-M is substantially outperformed by the RP model for some queries (5/15). For instance, for query Q21 ("russia meteor"), the difference is about 100%. From the analysis of the temporal distribution of the true relevant documents for this query, shown in Figure 8, we believe that the rationale for the low performance of our method is twofold. Firstly, it is likely that the query-terms alone return many documents that are uniformly distributed over a longer period of period. Thus, when we apply the rank fusion method and the density kernel function, it is difficult to keep only documents distributed over a shorter time period, which is mainly the first 50 hours as shown in Figure 8. Secondly, while the F-measure value may seem like a somewhat low, we feel that it is reasonable, given that even the RP model which is good in dealing for recency queries, fails in returning good results for this query (Q21). We note that if the topical model returns many non relevant results for our model and for the baselines because of sparsity, neither TTD-M nor RP will perform well after the reranking. This explains the zero F-measure value for query Q24 for both of the models (cf., Table IV).

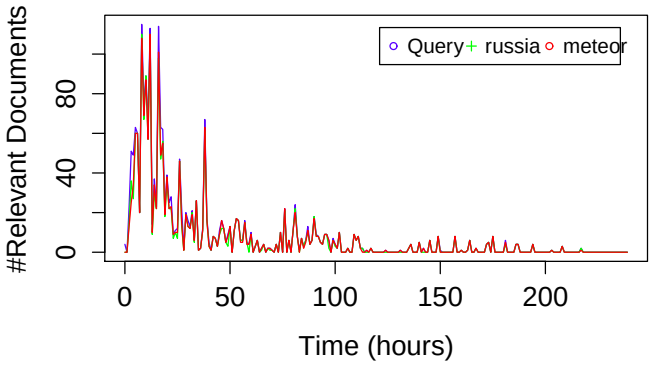


Fig. 8: Time-series of query Q21 and its within terms in the true relevant documents of the TREC 2014 TS track.

VI. Conclusion and Future Work

In this paper, we designed and evaluated a time-aware ranking approach aggregating the topical relevance matching criterion with a temporal relevance factor, based on a query term cross-correlation analysis. The intuition behind this approach is to boost documents that are published in the same time periods as a large number of relevant documents considering all of the terms forming the query. The analysis of the query-term temporal dependencies show a significant correlation of the temporal series among the terms belonging to the same query. Experiments using a large collection of news articles shows that the use of this property brings significant improvements and may be of benefit for temporal ranking. For future research, we plan to undertake large-scale experimental evaluation using other types of (temporal) collections in order to gauge the impact of collection-dependent parameters on the retrieval effectiveness (such as the temporal density kernel) and assess about the generalizability of our approach to other settings.

References

[1] R. Campos, G. Dias, A. M. Jorge, and A. Jatowt, "Survey of temporal information retrieval and related applications," *ACM Comput. Surv.*, vol. 47, no. 2, pp. 15:1–15:41, 2014.

[2] K. Berberich, S. Bedathur, O. Alonso, and G. Weikum, "A language modeling approach for temporal information needs," in *Proceedings of the 32Nd European Conference on Advances in Information Retrieval*, ser. ECIR'2010. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 13–25.

[3] D. Metzler, R. Jones, F. Peng, and R. Zhang, "Improving search relevance for implicitly temporal queries," in *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '09. New York, NY, USA: ACM, 2009, pp. 700–701.

[4] W. Dakka, L. Gravano, and P. G. Ipeirotis, "Answering general time-sensitive queries," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 2, pp. 220–235, 2012.

[5] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha, "Time is of the essence: Improving recency ranking using twitter data," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 331–340.

[6] M. Efron and G. Golovchinsky, "Estimation methods for ranking recent information," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '11. New York, NY, USA: ACM, 2011, pp. 495–504.

[7] K. Massoudi, M. Tsagkias, M. de Rijke, and W. Weerkamp, "Incorporating query expansion and quality indicators in searching microblog posts," in *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ser. ECIR'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 362–367.

[8] X. Li and W. B. Croft, "Time-based language models," in *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, ser. CIKM '03. New York, NY, USA: ACM, 2003, pp. 469–475.

[9] J. A. Aslam and M. Montague, "Models for metasearch," in *Proceedings of the 24th annual International* ACM SIGIR *Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2001, pp. 276–284.

[10] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '09. New York, NY, USA: ACM, 2009, pp. 758–759.

[11] O. Alonso, M. Gertz, and R. Baeza-Yates, "On the value of temporal information in information retrieval," *SIGIR Forum*, vol. 41, no. 2, pp. 35–41, 2007.

[12] S. Lin, P. Jin, X. Zhao, and L. Yue, "Exploiting temporal information in web search," *Expert Syst. Appl.*, vol. 41, no. 2, pp. 331–341, 2014.

[13] S. Nunes, C. Ribeiro, and G. David, "Use of temporal expressions in web search," in *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*, ser. ECIR'08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 580–584.

[14] D. Metzler, C. Cai, and E. Hovy, "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*, ser. NAACL HLT '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 646–655.

[15] M. E. Renda and U. Straccia, "Web metasearch: rank vs. score based rank aggregation methods," in *Proceedings of the 2003 ACM Symposium on Applied Computing*, ser. SAC '03. New York, NY, USA: ACM, 2003, pp. 841–846.

[16] L. Akritidis, D. Katsaros, and P. Bozanis, "Effective rank aggregation for metasearching," *Journal of System Software*, vol. 84, no. 1, pp. 130–143, 2011.

[17] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," in *Proceedings of the 10th International Conference on World Wide Web*. New York, NY, USA: ACM, 2001, pp. 613–622.

[18] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top k lists," in *Proceedings of the fourteenth annual* ACM-SIAM *symposium on Discrete algorithms*, Philadelphia, PA, USA, 2003, pp. 28–36.

[19] M. Farah and D. Vanderpooten, "An outranking approach for rank aggregation in information retrieval," in *Proceedings of the 30th annual International* ACM SIGIR *Conference on Research and Development in Information Retrieval*, ser. SIGIR '07. New York, NY, USA: ACM, 2007, pp. 591–598.

[20] M. Montague and J. A. Aslam, "Condorcet fusion for improved retrieval," in *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, ser. CIKM '02. New York, NY, USA: ACM, 2002, pp. 538–548.

[21] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to Time Series Analysis and Forecasting*, ser. Wiley Series in Probability and Statistics. New York, NY: Wiley, 2008.

[22] M. Efron, "Linear time series models for term weighting in information retrieval," *Journal of the American Society for Information Science and Technology*, pp. 1299–1312, 2010.

[23] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Trans. Inf. Syst.*, vol. 22, no. 2, pp. 179–214, Apr. 2004.