Temporal Diversification of Search Results

Klaus Berberich  
Max Planck Institute for Informatics  
Saarbrücken, Germany  
kberberi@mpi-inf.mpg.de

Srikanta Bedathur  
Indraprastha Institute of Information Technology  
New Delhi, India  
bedathur@iiitd.ac.in

ABSTRACT
We investigate the notion of temporal diversity, bringing together two recently active threads of research, namely temporal ranking and diversification of search results. A novel method is developed to determine search results consisting of documents that are relevant to the query and were published at diverse times of interest to the query. Preliminary experiments on twenty years’ worth of newspaper articles from The New York Times demonstrate characteristics of our method and compare it against two baselines.

Categories and Subject Descriptors
H.3.3 [Information Search & Retrieval]: Search process

Keywords
Time, Diversity, Temporal Information Retrieval

1. INTRODUCTION
Preservation and digitization efforts have resulted in more collections of born-digital or now-digital documents. These collections are large, consisting of millions or billions of documents, and often also longitudinal, spanning decades if not centuries. Concrete examples include web archives, newspaper archives, and collections of digitized books, but also the Web and social media, which have grown out of infancy and now contain documents from more than a decade ago.

For longitudinal document collections, time is an essential dimension and there has been a wealth of work during the past decade that tries to leverage temporal information (e.g., publication timestamps or temporal expressions) for tasks in Information Retrieval. For example, when ranking documents in response to a query, their associated temporal information can be used to systematically favor more recent documents published at time points of interest to the query [5], or documents whose content refers to a time period explicitly stated in the query [3]. Our focus in this work is on methods of the second kind, which target temporal queries [7] and seek to return relevant documents published at times of interest to the query at hand. None of the existing methods makes an effort to not only return relevant documents but also ensure that they were published at diverse times of interest to the query. Often, as we demonstrate in our experiments, a single time of interest to the query ends up dominating the returned documents, thus hampering the user experience. We argue that temporal diversity of search results is desirable when search longitudinal document collections, in particular for (a) temporally ambiguous queries (e.g., summer olympics and presidential election) that have multiple times of interest and (b) temporal informational queries (e.g., boeing, nuclear energy, clint eastwood) that lack any clear time of interest—still benefit from documents published at diverse times.

Diversification of search results has also been a recently active thread of research. The focus here is on ambiguous queries with different conceivable query intents behind them. Existing methods [2, 10, 11] determine results that provide a good mix of documents relevant to the different query intents. Query intents can either be modeled explicitly (e.g., as topics in a taxonomy) or implicitly (e.g., based on common reformulations of the query). No existing methods has considered time as a dimension for diversification.

In this work, we bring together these two threads of research to address the need for temporal diversity when searching longitudinal document collections. To this end, we make the following contributions: (i) a novel method for temporal diversification of search results, combining ideas from [5] and [2], (ii) preliminary experiments on The New York Times Annotated Corpus, investigating the characteristics of our method and comparing it against two baselines, (iii) a discussion of open issues and future work.

2. RELATED WORK
In this section, we present a brief overview of existing related work, which can be coarsely categorized as follows:

Temporal Information Retrieval. Closest to this work are efforts that leverage temporal information such as publication timestamps to improve the ranking of search results. Li and Croft [8] introduce a time-dependent prior in a language modeling approach to favor recent documents. Peetz et al. [9] compare several choices of the document prior motivated by findings from cognitive psychology. Other methods to favor recent documents in a language modeling approach such as time-dependent smoothing are explored by Efron and Golovchinsky [6]. Dakka et al. [5] describe a more gen-
eral approach that favors documents published around time points of interest to the query. To identify those, they build on Jones and Díaz [7] who describe features and methods to classify queries as atemporal, temporally unambiguous, or temporally ambiguous.

Novelty & Diversity. Carbonell and Goldstein [4] formulate the idea of maximal marginal relevance (MMR) in the context of summarization. Interest in the topic of diversity has been revived recently, stirred by the work of Agrawal et al. [2]. Their focus is on ambiguous queries (e.g., jaguar) with different conceivable query intents behind them (e.g., finding out about the car, animal, or operating system). Query intents are modeled explicitly as topics in a taxonomy according to which also documents are classified. The search result is then determined so as to maximize the probability that an average user sees at least one relevant document, which is formulated as a \( NP \)-hard but well-approximable optimization problem. Santos et al. [10], in follow-up research, consider how query intents can be modeled implicitly based on common reformulations of the query. Welch et al. [11] adapt the problem formulation for informational queries, for which users typically need to see more than one relevant document.

3. TEMPORAL DIVERSIFICATION

When aiming for temporal diversity, our objective is to determine a search result that consists of relevant documents published at diverse times of interest to the query. Thus, as a concrete example, for a temporally ambiguous query such as presidential election, returned documents should allude to an election of a president and have been published around diverse time points when such elections took place. In the following, we develop our method pursuing that objective.

Model. We assume a timestamped document collection \( \mathcal{D} \). Every document \( d^t \in \mathcal{D} \) carries a timestamp \( t \in \mathcal{T} \) conveying its publication time. The time domain \( \mathcal{T} \) is assumed to be discrete with timestamps indicating the number of time units (e.g., days) that have passed since a reference time (e.g., the UNIX epoch). We let \( \mathcal{Y} \) denote a vocabulary of searchable terms. Documents are multisets of terms drawn from this vocabulary. We use \( tf(w, d) \) to indicate how often the term \( w \) occurs in document \( d \) and let \( |d| \) denote the total number of term occurrences in the document. We use \( D \) to refer to a virtual document concatenating all documents from the collection. Likewise, \( D_t \) denotes a virtual document concatenating all documents published at time \( t \).

Temporal Ranking. We now describe, as a stepping stone, one of the methods from Dakka et al. [5]. The method is based on language modeling and ranks documents according to their probability

\[
P(d^t | q) = P(d, t | q) \propto P(d | q) \cdot P(t | q) \propto P(q | d) \cdot P(t | q) .
\]

The first term \( P(q | d) \) is the likelihood of generating the query \( q \) from document \( d \). It is estimated using Dirichlet smoothing as

\[
P(q | d) = \prod_{w \in q} \frac{tf(w, d) + \mu}{|d| + \mu} ,
\]

The second term \( P(t | q) \) conveys the relative importance of the time point \( t \) for the query \( q \) and is estimated as

\[
P(t | q) \propto \prod_{w \in q} P(w | t) = \prod_{w \in q} \frac{tf(w, D_t)}{|D_t|} .
\]

In practice, to account for effects such as delayed coverage of events in the media, this probability is smoothed using techniques from time-series analysis. Our concrete implementation smooths using a sliding window of width \( 2 \cdot w + 1 \) centered around \( t \). For the query presidential election, as our running example, Figure 2 plots the obtained values \( P(t | q) \).

Intuitively, this model aims to bring up documents published around times of interest to the query. However, as demonstrate in our experiments, the model does not ensure for a temporally ambiguous query like presidential election that all—or at least multiple—times of interest to the query are covered by the returned documents.

Temporal Diversity. To accomplish temporal diversity, we adapt the method proposed by Agrawal et al. [2] to our setting. We consider time points as query intents, as opposed to topics from a taxonomy in their case. One important difference is that topics are inherently categorical, whereas time points are numerical—a fact that our method has to take into account. While users who issue the query jaguar with cats in mind are certainly not interested in cars, users may tolerate or be indifferent to small differences in time. Thus, users interested in a particular time point may also be satisfied with documents published a few days earlier or later. As in the original method, our objective is to determine a sequence \( S = \{d_1^t, \ldots, d_n^t\} \) of documents that maximizes the probability that an average user sees at least one relevant document. This can be cast into the optimization problem

\[
\arg \max_S \sum_{t \in \mathcal{T}} P(t | q) \cdot \left( 1 - \prod_{d_i^t \in S} (1 - P(R(t, t_i) \cdot P(R(q, d_i))) \right) .
\]

Here, \( P(t | q) \) is defined as above and indicates the relative importance of the time point \( t \) for the query \( q \). Further, \( P(R(t, t_i)) \) indicates the probability that a user interested in time point \( t \) is satisfied with a document published at time \( t_i \). Similarly, \( P(R(q, d_i)) \) is the probability that a user issuing the query \( q \) is satisfied with the document \( d_i \). This models a user who judges the relevance of document \( d_i^t \) to query \( q \) and time point \( t \) by considering in isolation the content and publication timestamp of the document.

We model the probability that the time point \( t_i \) is relevant for a user interested in time point \( t \) using the sigmoid function

\[
P(R(t, t_i)) = \frac{1}{1 + e^{-w \cdot |t - t_i|}} ,
\]

so that the user is satisfied if the two time points \( t \) and \( t_i \) are less than \( w \) time units apart, but becomes quickly dissatisfied otherwise.

The probability \( P(R(q, d_i)) \) is estimated directly from the query likelihoods of the language model as

\[
P(R(q, d_i)) = \frac{P(q | d_i)}{\max_{d_j \in D} P(q | d_j)} .
\]

To determine a temporally diversified result \( S \), we use the greedy algorithm described in [2], which gives a \((1 - 1/e)\)-approximation guarantee.
4. EXPERIMENTS

We next describe our preliminary experimental evaluation. Its purpose is twofold: (i) see whether our method accomplishes temporal diversity and (ii) study its sensitivity to the parameter \( w \).

**Setup & Dataset.** We implemented all methods in a prototype system using Java. The New York Times Corpus [1], consisting of about 1.8 million newspaper articles published between 1987 and 2007, serves as a publicly-available longitudinal document collection for our experiments.

**Methods** that we compare are: (i) \( \text{LM} \) as the unigram language model with Dirichlet smoothing (according to Eq. 2), (ii) \( \text{LM+T} \) as the temporal language model by Dakka et al. [5] (described in Eq. 1), and (iii) our temporal diversification method \( \text{LM+T+D} \) (according to Eq. 4). We use \( \mu = 1000 \) for Dirichlet smoothing. When varying the value of \( w \), as indicated below, it is kept consistent across methods.

**Results.** Figures 1 (a)-(c) show top-5 documents returned by the methods under consideration for three different queries. For all of them, we set \( w = 180 \), which roughly models a user for whom time points no more than a year apart are indistinguishable. What can be observed across all three queries is that \( \text{LM} \) returns a top-5 result consisting of documents published at diverse points in time, which were not necessarily close to times of interest to the query. For the query presidential election, as a concrete example, only one of the documents was published in an actual election year. \( \text{LM+T} \) returns a top-5 result mostly consisting of documents published around what is arguably the most interesting time for the query, considering the time period 1987–2007 covered by the document collection, namely (a) the begin of the war in Iraq in March 2003, (b) Angela Merkel’s election in September 2005, and (c) the highly controversial U.S. presidential election in November 2000 (cf. Figure 2). Thus, for all queries, a single time of interest to the query ends up dominating the search result. When looking at the top-5 results produced by \( \text{LM+T+D} \), we see that it brings up documents published at diverse times of interest to the respective query. More specifically, documents are returned that cover (a) Operation Desert Storm and more recent developments, (b) Angela Merkel’s early career as well as her first term in office, and (c) four out of five presidential elections in the time period covered by the document collection.

Figure 1 (d) shows top-5 results obtained by \( \text{LM+T+D} \) for the query microsoft quarter earnings using different values of \( w \). When considering roughly a two-week window \( (w = 7) \), three out of five documents in the result are from the quarter in which Windows 95 was released. For \( w = 45 \), corresponding to a three-month interval, we see document from different quarters, three out of which are still from 1995. When setting \( w = 180 \), thus making the user indifferent to time differences of less than a year, documents cover earnings reports from five different years.

**Summary.** Although clearly anecdotal, the results presented indicate that (i) our method \( \text{LM+T+D} \) accomplishes temporal diversity in search results with (ii) a degree of temporal diversity that can be controlled using the parameter \( w \).

**5. OPEN ISSUES & FUTURE WORK**

This work constitutes only a first step, leaving several issues open and raising questions for future research.

**Open Issues:**

- What are ways to estimate \( P(R(t,t_i)) \) and \( P(R(q,d_i)) \) that are more principled than ours?
- How to design a comprehensive experimental evaluation? This includes the choice of document collection, queries, and effectiveness measures. While diversity has been addressed in the TREC Novelty Track, our setting does not permit gathering relevance judgments per query intent (i.e., time point) and the suitability of intent-aware effectiveness measures is unclear.
- How can the value of \( w \) be adjusted automatically? Possible clues may obtained from the values \( P(t|q) \), for instance, by examining the corresponding time series for bursts and periodicity.

**Future Work.** Our focus in this work has been on diversifying search results based on publication timestamps of documents. Alternatively, temporal diversification could be done based on temporal expressions contained in the documents, in order to return relevant documents that refer to diverse times of interest to the query. Another question for future work is how search results can be diversified along multiple orthogonal dimensions (e.g., topic and time).

6. REFERENCES

### (a) iraq war ($w = 180$)

<table>
<thead>
<tr>
<th>LM</th>
<th>LM+T</th>
<th>LM+T+D</th>
</tr>
</thead>
<tbody>
<tr>
<td>THE WORLD; This Spring</td>
<td><a href="http://bit.ly/1t1b7Ip">http://bit.ly/1t1b7Ip</a> @ 1988/04/10</td>
<td>Economic Scene; ...</td>
</tr>
</tbody>
</table>

### (b) angela merkel ($w = 180$)

<table>
<thead>
<tr>
<th>LM</th>
<th>LM+T</th>
<th>LM+T+D</th>
</tr>
</thead>
</table>

### (c) presidential election ($w = 180$)

<table>
<thead>
<tr>
<th>LM+T+D ($w = 7$)</th>
<th>LM+T+D ($w = 45$)</th>
<th>LM+T+D ($w = 180$)</th>
</tr>
</thead>
</table>

### (d) microsoft quarter earnings

<table>
<thead>
<tr>
<th>LM+T+D ($w = 7$)</th>
<th>LM+T+D ($w = 45$)</th>
<th>LM+T+D ($w = 180$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dow Rises 11.56 as Technology Stocks ...</td>
<td><a href="http://bit.ly/171O63y">http://bit.ly/171O63y</a> @ 1995/10/18</td>
<td>Microsoft Manages to Post ...</td>
</tr>
<tr>
<td>Microsoft Net Up 10.2% ...</td>
<td><a href="http://bit.ly/1b2qylQ">http://bit.ly/1b2qylQ</a> @ 1989/04/20</td>
<td>Microsoft Net Up 10.2% ...</td>
</tr>
<tr>
<td>Microsoft Manages to Post ...</td>
<td><a href="http://bit.ly/1N4h6yQ">http://bit.ly/1N4h6yQ</a> @ 1995/04/14</td>
<td>Microsoft Has Strong Profit For Quarter ...</td>
</tr>
</tbody>
</table>

Figure 1: Top-5 Results