Time Will Tell: Leveraging Temporal Expressions in IR*

Irem Arikan Srikanta Bedathur Klaus Berberich

Max-Planck Institute for Informatics
Saarbrücken, Germany
{iarikan, bedathur, kberberi}@mpi-inf.mpg.de

ABSTRACT

Temporal expressions, such as *between 1992 and 2000*, are frequent across many kinds of documents. Text retrieval, though, treats them as common terms, thus ignoring their inherent semantics. For queries with a strong temporal component, such as U.S. president *1997*, this leads to a decrease in retrieval effectiveness, since relevant documents (e.g., a biography of Bill Clinton containing the aforementioned temporal expression) can not be reliably matched to the query.

We propose a novel approach, based on language models, to make temporal expressions first-class citizens of the retrieval model. In addition, we present experiments that show actual improvements in retrieval effectiveness.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithms, Experimentation, Performance

Keywords

Temporal Information Retrieval, Language modeling

1. INTRODUCTION

Increasing amounts of content, not only created at different times but also pertaining to different times, are available on the World Wide Web. Prominent examples of such content include news articles, blogs, and wikis. Typical approaches to retrieval either treat the temporal expressions contained in these documents simply as common terms, or take the creation time of a document as a surrogate for the temporal context of the document's content.

However, both families of approaches fail to capture the semantics inherent to the time dimension. Treating temporal expressions as common terms, on the one hand, ignores their inherent semantics. Unless document and query contain exactly the same temporal expression, the document will not be ranked high in the results. The creation time of a document, on the other hand, can be way off the time the contents of the document pertain to – think of a web page describing a futuristic science-fiction plot or a Wikipedia article about the French revolution.

Search Engine Y

1. List of state leaders in 1977
http://en.wikipedia.org/wiki/List_of_state_leaders_in_1977
2. Prime minister
http://en.wikipedia.org/wiki/Prime_minister
3. List of state leaders in 1976
en.wikipedia.org/wiki/List_of_state_leaders_in_1976
4. List of state leaders in 1974
http://en.wikipedia.org/wiki/List_of_state_leaders_in_1974
5. List of state leaders in 1978
http://en.wikipedia.org/wiki/List_of_state_leaders_in_1978

Search Engine G

1. List of state leaders in 1977
http://en.wikipedia.org/wiki/List_of_state_leaders_in_1977
2. French municipal elections, 1977
http://en.wikipedia.org/wiki/French_municipal_elections,_1977
3. France-Albert René
http://en.wikipedia.org/wiki/France-Albert_René
4. 1977
http://en.wikipedia.org/wiki/1977
5. Anthony Eden
http://en.wikipedia.org/wiki/Anthony_Eden

Figure 1: Search results "Prime Minister France 1977"

As a consequence, retrieval effectiveness suffers for queries that have a strong temporal component (e.g., such aimed at finding historical information). To illustrate this problem, consider a user who wants to find out who was prime minister of France in 1977. We ran the query "Prime Minister France 1977" on two popular web search engines – codenamed Y and G – while restricting the domain of search to `http://en.wikipedia.org/`. The top-5 answers for the query are listed in Figure 1. As these results show, the top result for the query is simply the full list of world leaders in 1977 – a special feature of Wikipedia that is typically not available in text collections. When ignoring this special result, none of the remaining results is relevant to our information need.

In order to improve retrieval effectiveness for such queries, it is therefore essential to pay special attention to temporal expressions contained in documents. In this paper, we address this very issue. Our key contribution is a novel approach that seamlessly integrates the temporal dimension into a language model based retrieval framework. Experimental evidence shows that our approach yields improvements in retrieval effectiveness. For instance, when evaluating the above query Prime Minister France *1977* using our approach, we obtain the article about Raymond Barre (http://en.wikipedia.org/Raymond_Barre), who was prime minister of France at the time of interest, at the second position.

*Partially supported by the EU within the 7th Framework Programme under contract 216267 "Living Web Archives (LiWA)"

WSDM'09, February 9–12, 2009, Barcelona, Spain.

2. MODEL

In this section, we lay out the model and the notation that will be used throughout the remainder. We let D denote our *document collection*. When modeling the contents of a document $d \in D$, we distinguish between *terms* and *temporal expressions*. Formally, a document consists of a bag of textual terms d_{tx} and a bag of temporal expressions d_{te}. A temporal expression T found in a document is a time interval $T = [b, e]$ with a begin boundary b and end boundary e drawn from a time domain. Queries in our setting consist of a textual part q_{tx} and temporal part q_{te}. The textual part q_{tx} is a set of terms and can thus be thought of as a standard keyword query. Analogously, the temporal part q_{te} is a set of temporal expressions that captures the times of interest to the user. As an example, a user interested in who were presidents of the U.S. in the 1990s could formulate the query U.S. president *1990s*.

3. LEVERAGING TEMPORAL EXPRESSIONS

We now proceed to the core of this work and describe how temporal expressions can be leveraged to improve retrieval effectiveness.

3.1 Ponte and Croft's Model

Our approach builds on *language models* as originally proposed by Ponte and Croft [13]. Due to space constraints we only give an informal description of their approach and point to the original work [13] for full details. For a recent more complete description of language models, we refer to to Manning et al. [9].

In Ponte and Croft's approach each document has a generative model of terms associated. The probability $P(t|d_{tx})$ of producing the term t from document d_{tx} depends on the term frequency of t in d_{tx}, but also on the collection frequency of t (i.e., its total number of occurrences in D).

Assuming independence for the generation of individual terms, the relevance of document d_{tx} to the query q_{tx} is then assessed as the probability of generating q_{tx} from the generative model associated with d_{tx}, i.e.,

$$P(q_{tx}|d_{tx}) = \prod_{q \in q_{tx}} P(q|d_{tx}) \times \prod_{q \notin q_{tx}} 1.0 - P(q|d_{tx}) \,. \quad (1)$$

In the remainder, we will refer to the Ponte and Croft model simply as LM.

3.2 Filtering Model

As we argued in the introduction, treating temporal expressions as standard terms is treacherous, as their inherent semantics is lost. For our earlier query example U.S. President *1990s*, a document mentioning that Bill Clinton was president between 1992 and 2000 would be treated equal to a document mentioning the president but not containing any temporal expression. Likewise, if a user is interested in what happened in San Francisco on *April 18th, 1906*, a document talking about a severe earthquake that happened in the 1900s would not be given preference to any other document mentioning San Francisco. As a third and final example, consider a user interested in British Punk Rock *between 1975 and 1980*. An article about The Clash stating that the famous punk band's active period was between 1976 and 1986 would not be favored, since the years stated in the document mismatch the years in the user's query.

These three examples are representative of different cases, namely those where the document contains a temporal expression that (i) is a *superinterval*, (ii) a *subinterval*, (iii) or an *overlapping interval* of a temporal expression given by the user. Intuitively, a document containing temporal expressions of these kinds is favorable to documents that do not contain any relevant temporal expressions.

Our first approach to take into account temporal expressions follows this intuition in a radical way. The idea behind the approach, coined LMF, is to not report documents that do not contain any temporal expressions of relevance to the user. The approach therefore filters out documents that do not contain (i) a superinterval, (ii) a subinterval, or (iii) an overlapping interval of a temporal expression specified in the user's query.

Formally, LMF reports only documents from the query-dependent subset of the collection

$$D(q_{te}) = \{\, d \in D \mid \exists\, T \in q_{te}\ \exists\, T' \in d_{te} \ :\ T \cap T' \neq \emptyset \} \,. \quad (2)$$

The relative ranking of result documents is exactly the same as the one obtained from the Ponte and Croft model for the textual part q_{tx} of the query. In fact, the approach is not dependent on the use of language models, but can be used with other relevance models as, for instance, Okapi BM25 [14]. Moreover, it can easily be implemented on top of an existing system as a post-filtering step.

3.3 Weighted Model

One drawback of the LMF approach just described is that it assumes a black-and-white perspective on the world. A document is either considered or not – there is no thing in between. In particular, the approach does not take into account (i) how many relevant temporal expressions a document contains and (ii) how closely they match the temporal expressions specified in the user's query. With regard to the second point and considering our earlier example, consider two documents that talk about earthquakes in San Francisco in *April 1906* and the *1900s*, respectively. Given otherwise equal relevance of the two documents, it is reasonable to favor the first document, as the temporal expression contained is closer to the temporal expression specified in the user's query, namely, *April 18th, 1906*.

Our second approach, coined LMW addresses these issues. LMW assigns higher relevance to a document, if it contains more temporal expressions that provide a closer match to the temporal part q_{te} of the user's query.

At the core of LMW lies a generative model for temporal expressions. Using this, we estimate the probability of generating the temporal part q_{te} of the user's query from the temporal expressions contained in a document. We assume that the generation of the textual query part q_{tx} and the temporal query part q_{te} happen independently, giving us

$$P(q|d) = P(q_{tx}|d_{tx}) \ \times\ P(q_{te}|d_{te}) \,, \quad (3)$$

where $P(q_{tx}|d_{tx})$ is estimated based on the Ponte and Croft model described above. In analogy to their model, we assume that temporal expression in q_{te} are generated independently. Therefore, the probability of generating the temporal query part q_{te} from the document d_{te} is

$$P(q_{te}|d_{te}) = \prod_{Q \in q_{te}} P(Q|d_{te}) \,. \quad (4)$$

We now introduce a generative model that determines the probability $P(Q|d_{te})$ of producing the temporal expression Q from the document d_{te}. In a first step, we draw a single temporal expression $T = [b, e]$ at uniform from all the temporal expressions contained in the document d_{te}. In a second step, we estimate the probability of generating the temporal expression $Q = [b', e']$ given T. Putting these two

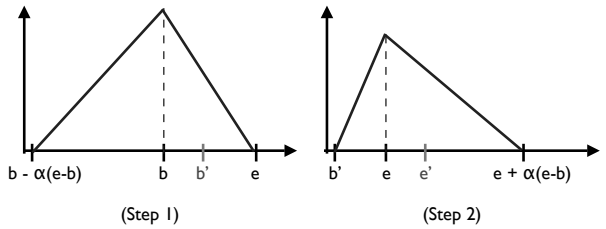

Figure 2: Generating $Q = [b', e']$ from $T = [b, e]$

steps together, we yield

$$P(Q|d_{te}) = \frac{1}{|d_{te}|} \cdot \sum_{T \in d_{te}} P(Q|T) \; . \tag{5}$$

The generative model associated with T must meet the desiderata motivated above. First, it should only produce temporal expressions whose time interval intersects with T. Second, because we favor documents that contain temporal expressions close to a temporal expression Q specified in the query, if Q is not close to T, it should have low probability of being produced by the generative model associated with T.

We subdivide the process of generating Q from T into first choosing the begin boundary b' and then the end boundary e'. For the probability of generating Q from T, we thus write

$$P(Q|T) = P(b') \cdot P(e'|b') \; . \tag{6}$$

Figure 2 provides a visualization of the generative process and the underlying probability distributions. The choice of b' is constrained as $b' \leq e$, since we must not generate an interval that does not intersect with T. Further, since we want temporal expressions close to T to have a higher probability of being generated, values of b' close to b should be favored. We employ a *triangular distribution* [7] to meet these criteria. The triangular distribution is defined by three parameters $x \leq y \leq z$ with $[x, z]$ being the support interval where the distribution assigns non-zero probabilities and y being the point having maximal probability. On the intervals $[x, y]$ and $[y, z]$ assigned probabilities are linearly increasing and decreasing, respectively. For determining $P(b')$, we employ a triangular distribution having parameters

$$x = b - \alpha \cdot (e - b) \quad y = b \quad z = e \; ,$$

where $\alpha \geq 0$ is a tunable parameter that constrains the choice of b' taking into account $(e - b)$, i.e., the length of T.

Having chosen b', the end boundary e' remains to be picked. This choice is constrained by $b' \leq e'$. Following the same reasoning as above, we again employ a triangular distribution having parameters

$$x = b' \quad y = e \quad z = e + \alpha \cdot (e - b)$$

to determine $P(e'|b')$. Again, the parameter α constrains the choice of e' depending on the time spanned by T.

4. PRELIMINARY EXPERIMENTS

To validate our hypothesis that temporal expressions can help to improve retrieval effectiveness, we conducted a preliminary series of experiments. Our finding obtained from it are the subject of this section.

Setup. We implemented the proposed methods using the Terrier [11] platform and, in particular, their implementation of the Ponte and Croft model. For the extraction of temporal expressions, we follow a simplified version of the approach described in Zhang et al. [15]. Each document is matched against a set of regular expression capturing common formats of temporal expressions, for instance, `[dD]uring (\d{1,4})( BC| B.C.){0,1}`. The extracted temporal expressions are mapped to their corresponding time intervals, which are then stored in a MySQL database. In our experiments, we consider the three approaches described in Section 3 – we do not compare against Y and G that were mentioned in the introduction, since we can neither ensure the use of the same dataset, nor do we possess enough knowledge about their internals. For LMW we used value of $\alpha = 3.0$ to produce the results – for the queries presented we found results to be fairly robust across different choices of α. Notice that for the Ponte and Croft language model the query is processed as specified by the user, i.e., no temporal expressions are extracted. As a concrete example, the query Earthquake *1980 – 1990* is sent to LM as $q_{tx} = \{\text{Earthquake}, 1980, 1990\}$ but to the other two methods as

$$q_{tx} = \{\text{Earthquake}\} \quad q_{te} = \{[1980, 1990]\} \; .$$

Dataset. As a dataset we use a snapshot of the English Wikipedia [3] taken in early May 2007. This dataset contains about 2M encyclopedia articles as HTML pages.

Figure 3 shows the titles of the five highest-ranked result documents for different queries. For the first query Prime Minister France *1977* (Figure 3(a)), it can be seen that only LMW brings up the encyclopedia article for Raymond Barre who was prime minister at the time of interest. For our second query Spanish Painter *18th Century* (Figure 3(b)), LMF and LMW have two and three Spanish painters from the period of interest among their results, respectively. In contrast, results from LM are fairly broad and do not contain any person-specific articles. For the query Sea Battle *1650-1670* (Figure 3(c)) we observe that LM brings a good result to the top, but other than that produces only rather broad or non-relevant results. LMF brings up two results relating to sea battles that took place in the period of interest. The five highest-ranked results by LMW, finally, all relate to specific sea battles in the period of interest. A similar observation can be made for the query Earthquake *1980-1990* (Figure 3(d)). Three of the results produced by LMW are specifically dedicated to earthquakes that happened between 1980 and 1990. Results from LMF, in contrast, are fairly general. LM brings up two result relating to earthquakes at the time, but also two sports-related results.

Summary. The anecdotal results presented strongly indicate that LMF and LMW outperform the baseline, producing results that are both textually and temporally relevant. Comparing the two methods, it can be observed that LMW produced consistently better results than LMF.

5. RELATED WORK

We now briefly put this work in context with existing prior research. Alonso et al. [4] highlight the importance of temporal information for Information Retrieval and give an overview of existing approaches – the problem addressed in our work is explicitly mentioned as one not satisfactorily supported by today's search engines. Nunes et al. [10] discovered that, on average, temporal expressions are present in about 1.5% of web queries – queries about News and Sports were found to exhibit a significantly higher percentage.

Li and Croft [8] proposed time-based language models that take into account the publication times of documents as to favor, for instance, more recent documents. Del Corso et al. [6] studied the problem of ranking news articles, also

	LM	LMF	LMW
1	List of State Leaders in 1977	List of State Leaders in 1974	List of State Leaders in 1977
2	List of State Leaders in 1974	Antoine Pinay	**Raymond Barre**
3	List of State Leaders in 1976	Henri Queuille	Deputy Prime Minister of Canada
4	List of State Leaders in 1978	List of State Leaders in 1977	List of State Leaders in 1978
5	List of State Leaders in 1979	List of State Leaders in 2000	Minister of Territorial Development

(a) Prime Minister France *1977*

	LM	LMF	LMW
1	Art in Puerto Rico	**José del Castillo**	**José del Castillo**
2	Spanish Art	List of Spanish Artists	Roybal
3	Palazzo Bianco (Genoa)	Roybal	**Augustine Esteve**
4	Caprichos	**Augustine Esteve**	Maldonado
5	Portrait Painting	Francisco Eduardo Tresguerras	**Luis Egidio Meléndez**

(b) Spanish Painter *18th Century*

	LM	LMF	LMW
1	**Battle of Dunbar (1650)**	List of Norwegian Battles	**Battle of the Gabbard**
2	Monte Mataiur	**Battle of Portland**	**Battle of Portland**
3	St. George's Caye	Action of 22 February 1812	**Battle of Scheveningen**
4	Culrain Scottland	Naval Strategy	**Battle of the Kentish Knock**
5	First Anglo-Dutch War	**Battle of the Gabbard**	**Battle of Dungeness**

(c) Sea Battle *1650 – 1670*

	LM	LMF	LMW
1	San Jose Earthquakes	Earthquake	**1990 Luzon Earthquake**
2	Earthquake (Comics)	Intraplate Earthquake	Earthquake (Comics)
3	Joe Morrone Jr.	Earthquake Prediction	**1987 Edgecumbe Earthquake**
4	**Sant'Angelo dei Lombardi**	Parkfied Earthquake	**1985 Mendoza Earthquake**
5	**Clayton-Marsh Creek-Greenville Fault**	New Madrid Earthquake	Gap Hypothesis

(d) Earthquake *1980 – 1990*

Figure 3: Anecdotal query results

taking into account their time of publication and linkage among the articles. None of the approaches, however, considers temporal expressions contained in the documents.

In Pasça [12] temporal expressions are used to improve the performance of Question Answering for time-related questions, such as "When was the Taj Mahal built?". Answers are then obtained by aggregating over matching pieces of information and their contained temporal expressions.

The work closest to ours is Baeza-Yates [5] whose aim is to search information that refers to the future. The proposed retrieval model is focused on confidences associated with statements about the future, thus favoring relevant documents that are confident about their predictions regarding the future time of interest. The frequency of temporal expressions and their closeness to what the user is interested in, though, is not explicitly considered.

Finally, several prototypes are available that make use of temporal expressions when searching the Web, most notably, Google's Timeline View [1] and TimeSearch [2]. Details about their internals have not been published.

6. SUMMARY

Documents are often rich of temporal expressions whose inherent semantics is typically ignored by relevance models. As an effect, retrieval effectiveness suffers for information needs that have a strong temporal component. In this work, we proposed two methods to address this problem and presented first experimental evidence demonstrating their improving retrieval effectiveness.

Ongoing & Future Work. At the time of writing we are conducting further experiments that include additional data sets and involve real users to obtain relevance judgments. There are several interesting avenues of future research – incorporating information about the proximity between temporal expressions and regular terms is one of them.

7. REFERENCES

[1] Google's Timeline View `http://www.google.com/experimental/`.

[2] TimeSearch History`http://www.timesearch.info`.

[3] Wikipedia `http://www.wikipedia.org`.

[4] O. Alonso, M. Gertz, and R. Baeza-Yates. On the value of temporal information in information retrieval. *SIGIR Forum*, 41(2), 2007.

[5] R. A. Baeza-Yates. Searching the future. In *ACM SIGIR Workshop MF/IR*, 2005.

[6] G. M. D. Corso, A. Gulli, and F. Romani. Ranking a stream of news. In *WWW*, 2005.

[7] M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions 3rd ed.* Wiley, New York, 2000.

[8] X. Li and W. B. Croft. Time-based language models. In *CIKM*, 2003.

[9] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[10] S. Nunes, C. Ribeiro, and G. David. Use of Temporal Expressions in Web Search. In *ECIR*, 2008.

[11] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *OSIR*, 2006.

[12] M. Pasça. Towards temporal web search. In *ACM SAC*, 2008.

[13] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR*, 1998.

[14] S. E. Robertson and S. Walker. Okapi/keenbow at trec-8. 1999.

[15] Q. Zhang, F. M. Suchanek, L. Yue, and G. Weikum. TOB: Timely Ontologies for Business Relations. In *WebDB*, 2008.