Metric Spaces for Temporal Information Retrieval

Matteo Brucato[1] and Danilo Montesi[2]

[1] School of Computer Science, University of Massachusetts, Amherst, MA, USA
matteo@cs.umass.edu
[2] Department of Computer Science and Engineering, University of Bologna, Italy
montesi@cs.unibo.it

Abstract. Documents and queries are rich in temporal features, both at the meta-level and at the content-level. We exploit this information to define *temporal scope similarities* between documents and queries in *metric spaces*. Our experiments show that the proposed metrics can be very effective for modeling the relevance for different search tasks, and provide insights into an inherent asymmetry in temporal query semantics. Moreover, we propose a simple ranking model that combines the temporal scope similarity with traditional keyword similarities. We experimentally show that it is not worse than traditional keyword-based rankings for non-temporal queries, and that it improves the overall effectiveness for time-based queries.

1 Introduction

The amount of available digital information is constantly increasing—a phenomenon that many refer to as *big data*. As more and more information becomes available year after year, its variety and richness in terms of *temporal aspects* become more manifest. A recent research area aimed at incorporating temporal aspects in modern information retrieval systems is *temporal information retrieval* (TIR) [1]. By its literature, it is clear how time comes into play in many different facets and forms. For instance, what kind of temporal features should we consider? How can we define the "temporal needs" of users, and the "temporal intent" of queries? Some of these issues have been explored by the research community [2,3,4], but further works are still needed as we are far from having widely-accepted solutions.

In this work, we explore ways to improve traditional ranking models by considering what we call the *temporal scopes* of documents and queries. These indicate which periods of time the documents are about and which periods of time users are interested in when issuing time-based queries. For example, news stories very often refer to periods of time close to the publication dates, chapters from history books may refer to any past period, and tweets from the Twitter social network might conversely have very narrow time scopes.

There are quite a number of search tasks in which the temporal scope is of great importance. For instance, imagine an expert user (e.g. a librarian, a historian, or a philologist) who is searching a digital library or a digital historical

M. de Rijke et al. (Eds.): ECIR 2014, LNCS 8416, pp. 385–397, 2014.

archive to find information about a specific period of time. To be able to only filter documents by their creation date would be too limiting. So would be treating temporal expressions such as "last year" as simple search terms, since they indicate specific time periods and their meaning changes with time.

In this paper, we propose a new model for temporal information retrieval based on metric spaces. We study temporal aspects of documents and queries, and we use temporal expressions [5] extracted from their texts to model their temporal scope. References to temporal information are represented in the temporal domain as time intervals, and temporal similarities between documents and queries are defined according to them. We evaluate the effectiveness of our model through a series of experiments, whose results are twofold. First, they confirm that exploiting temporal information can enhance the effectiveness of traditional keyword-only models for temporal queries. Second, they provide insights into an inherent asymmetry in temporal query semantics, confirming our intuition which led us to the definition of generalized metrics for modeling the temporal relevance for different search tasks.

2 Related Work

The time dimension has been extensively studied in temporal databases [6]. More recently, its importance has also been acknowledged in information retrieval [1]. While earlier works mostly concentrated on exploiting temporal meta-information (such as the creation date of documents) [7], there is a more recent interest in considering temporal expressions extracted from the text to improve the effectiveness of ranking algorithms [8]. Recent advances in natural language processing (NLP) have made it possible to effectively identify and interpret these expressions in a variety of texts (see [9] for an overview of the problem). Nowadays, there are ready-to-use tools for extracting [10] and normalizing (i.e. interpreting) [11] them easily and reliably.

There are several aspects of temporal information retrieval that current research has been focusing on. The *temporal intent* of textual queries has been discussed, for instance, in [2]. The *implicit time* of textual search queries has been studied in [12,3], among others. First attempts to linearly combine non-temporal scores with temporal ones have been presented in [7,13]. Moreover, several workshops on time-related aspects of information retrieval have been recently organized, such as [14] and [15], thus showing the attention given to this topic by current research.

More importantly, there exists previous work aimed at showing how to improve the effectiveness of search engines on temporal queries. For instance, in [4] and [7] the authors presented different language models to address different temporal information needs. Our work differs from all existing works because it introduces the first non-probabilistic ranking model for the temporal scope of documents and queries in temporal information retrieval based on (generalized) metric spaces.

3 Motivation

Time is an important aspect in every collection of documents. It is a ubiquitous dimension that can be interrelated with all sorts of information. For instance, it can be attributed to events ("when" they take place) and facts ("when" they are true or false). It is so intrinsic that it is often taken for granted and, thus, disregarded.

At the **meta-level**, a document has a *creation date*, a *publication date*, a *revision date*, and so on. But a closer look at its **content-level** can reveal information about which times the document is about: Which *periods of time* are explicitly mentioned in the text? When did the *events* mentioned in the text happen? Searching collections at the meta-level is important, and it has been studied in several recent works [16]. In this work, we concentrate on the temporal information present at the content-level. This type of approach requires further steps in the acquisition process of the information contained in the documents. It requires tools to identify and interpret natural language expressions with temporal intent, which are called *temporal expressions* (or *timexes*). More precisely, by looking at the content-level, we are able to make sense of what we call the **temporal scope** of a document. This level of understanding is crucial when we want to search documents based on their temporal content.

Not surprisingly, queries are also rich in temporal features. At the **meta-level** we can identify, for instance, the issue date of a query (which is crucial for query log analysis like Google Trends[1]). But looking at the **content-level** (i.e. the text of queries) can reveal very specific query intents. For example, querying "Obama elections 2008" is very different from querying "Obama elections", for its intent of discriminating among events in a case of clear ambiguity. And the expression "last year" in the query "last year best movies" can be interpreted as a signal for the intent of drifting away from the plausible default behavior of always retrieving the latest information. Temporal expressions are present in a substantial fraction of queries (about 1.9%, as per [5]). Although this is not a very large percentage, when a query contains such a signal, its intent becomes very different from the default one.

As a concrete example, consider a user searching a digital library, an archive, or a crawled portion of the Web. She uses natural language queries to specify her information needs, which pertain not only to the textual content, but also to the temporal scope of the information content. For instance, she queries the phrase "balkan conflicts in 1912 and 1913" to retrieve relevant information about the Balkan Wars. Any information related to those events is relevant to the user, including the causes of the wars and the aftermath. In this scenario, the relevance of a document can be modeled by: (1) a traditional notion of keyword-based similarity (e.g. cosine similarity in the vector space model [17]); (2) a notion of *temporal scope similarity*, which might favor documents regarding time periods that are "close" to the query time period.

[1] `http://trends.google.com`

4 Temporal Scope Similarity Model

One of our major aims is to provide more evidence that a similarity measure based on the temporal scope of documents and queries can lead to improvements in the effectiveness of traditional ranking algorithms based solely on term statistics. With keyword-based similarity models, we can readily identify documents that are textually similar to the query, but we cannot easily distinguish between two documents that are textually too similar to one other. In the search space dictated by keyword similarity, all textually similar documents have similar representations, which means that these models are not rich enough for distinguishing among them. The time dimension, in some cases, can make it possible to have a clearer distinction. Moreover, similarity models based solely on term occurrences are not rich enough for capturing specific time-related aspects of queries and documents. In particular, we identify the following three characteristics of temporal expressions that cannot be modeled with simple keyword-based models, and that can lead to poor results.

Temporal Synonymy. Suppose we treat temporal expressions as in keyword-based models. That is, the expression "2014" is for us a simple term that can occur in a document or a query with a certain frequency. In this model, we would not be able to account for different ways of referring to the year 2014. But there are many. For instance, we could write "last year", or "next year", or even "in a decade", depending on the context.

Temporal Polysemy. Additionally, some temporal expressions can potentially refer to more than just one period of time. Consider for instance the expressions "every Tuesday", "yearly", or the implicitly temporal expression "super bowl".

Structured Domain. Periods of time can be modeled as intervals of numbers, i.e. they can be represented as *time* or *temporal intervals* (as it has been proposed in the context of temporal databases [6]). In the domain of temporal intervals, it is easier to define notions of *overlap*, *containment* and *distance* between them at the semantic level, rather than at the syntactic level.

We frame our model for temporal scope similarity as follows, embedding a notion of distance between time intervals:

1. Different temporal expressions can be mapped to the same temporal interval.
2. A single temporal expression can be mapped to multiple temporal intervals.
3. The temporal space is a metric space (Δ, δ), where δ is a distance function on Δ, modeling a notion of distance between temporal intervals.
4. The temporal scope of documents and queries are subsets of Δ, i.e., $T_D \in \mathcal{P}(\Delta)$ and $T_Q \in \mathcal{P}(\Delta)$, where D and Q are a document and a query, respectively.
5. Let δ^* be a distance function on $\mathcal{P}(\Delta)$ defined in terms of δ. Then, $(\mathcal{P}(\Delta), \delta^*)$ is a metric space for document and query representations, where δ^* models a notion of distance between them in terms of their temporal scope.

The temporal scope similarity can be defined as a similarity in the metric space $(\mathcal{P}(\Delta), \delta^*)$. We formalize these concepts in the next two sections.

5 Modeling the Temporal Scope

In this section, we formally define the domain of temporal intervals Δ, used to represent documents and queries. Our goal is to utilize temporal expressions extracted from the text to model the temporal scope. Temporal expressions are used in natural language texts to express temporality [8,4]. For instance, they might be used to state when a certain event happened (e.g. "two years ago Obama won the elections").

We model the temporal scope by mapping the temporal expressions to a *temporal domain* Δ, which is the set of all possible temporal intervals, represented as ordered pairs of integers. By doing so, we obtain a *temporal scope representation* for each document and each query, that we indicate with T_D and T_Q respectively, and such that $T_D \subseteq \Delta$ and $T_Q \subseteq \Delta$.

Definition 1. Chronon. *A* chronon *is the smallest discrete unit of time, i.e., an atomic time. It describes the granularity of the model. Examples of chronons are seconds, days, years, etc.*

Definition 2. Timeline. *Let* $t_{min}, t_{max} \in \mathbb{Z} : t_{min} \leq t_{max}$. *The* timeline *is the totally ordered set of numbers*

$$\Gamma = \{t \in \mathbb{Z} \mid t_{min} \leq t \leq t_{max}\}$$

in which each number corresponds to a different chronon, and consecutive numbers correspond to consecutive chronons. Therefore, t_{min} *and* t_{max} *correspond, respectively, to the first and the last chronons that can be captured by the timeline, and* $|\Gamma| = t_{max} - t_{min} + 1$ *is the cardinality of the timeline.*

Definition 3. Temporal Domain. *The* temporal domain *is the set*

$$\Delta = \{[s,t] \mid s,t \in \Gamma \text{ and } s \leq t\} \subseteq \Gamma \times \Gamma$$

that is, the set containing all pairs of timeline elements, internally ordered. It follows that the cardinality of Δ *is* $|\Delta| = \frac{|\Gamma|(|\Gamma|+1)}{2}$.

Definition 4. Temporal Intervals. *Let* Timex *be the set of all temporal expressions that can extracted either from documents or queries, and let* Ψ : Timex $\rightarrow$ $\mathcal{P}(\Delta)$ *be a function that maps temporal expressions to temporal intervals. Let* $e \in$ Timex *be a temporal expression. The set* $\Psi(e)$ *is the set of all the* temporal intervals *of the expression* e.

For convenience, we will also use Timex_Q and Timex_D to denote the set of expressions extracted from a query Q and a document D, respectively. Further, we will use $[s,t]_Q$ and $[s,t]_D$ to indicate whether the temporal interval $[s,t]$ has been extracted from the query Q or the document D, respectively. Notice that, at this point, the first two points of Sect. 4 are both satisfied.

Definition 5. Document/Query Temporal Scopes. *The* document temporal scope *and the* query temporal scope *are the document and the query representations in the temporal domain:*

$$T_D = \{[s,t]_D\} = \{[s,t] \in \Psi(e) \mid e \in \textsc{Timex}_D\} \subseteq \Delta \tag{1}$$

And similarly for T_Q. Notice that now point 4 of Sect. 4 is also satisfied.

6 Temporal Scope Similarity

To understand the difficulty of modeling a temporal similarity metric, consider the following example. Imagine two textually similar documents, one containing the time expression "during the twentieth century", and one containing the time expression "June 1950", as shown in Fig. 1. The temporal scope of the first document is broad, whereas the second one is narrow. Now, suppose a user formulates the query "between 1940 and 1960", as also shown in the picture. Which of the two documents would the user consider more relevant?

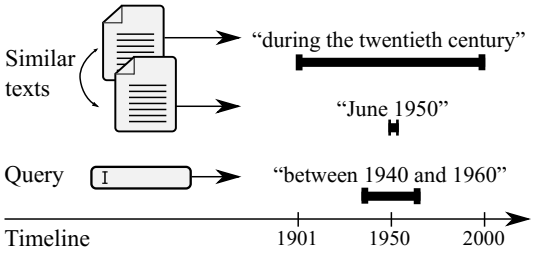


Fig. 1. Example of two textually similar documents with different temporal scopes, and a time-based query

This question cannot be answered without knowing the query semantics and how the user expressed her information needs as a textual query. Some users might consider the broader document more relevant because its temporal scope covers the query scope. Others might think that a broader document is less relevant because it is too generic, and that the narrower document is more relevant because it falls inside the query scope. Perhaps, some other users might think that the best document should have a temporal scope that matches exactly the query scope. Therefore, we propose to use three different *generalized metric spaces* (i.e., metric spaces in which some of the metric properties, in particular symmetry and coincidence, are relaxed) to capture these alternatives.

6.1 Generalized Metric Spaces

The goal of this section is to model δ^*. Since documents and queries are represented as sets of temporal intervals, i.e. T_D and T_Q respectively, as in (1),

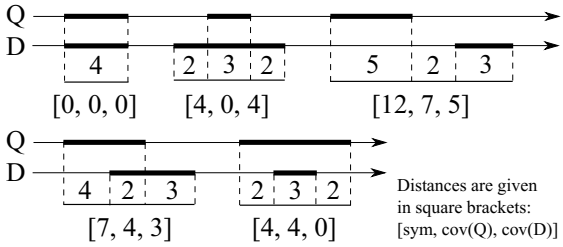


Fig. 2. Five different examples of query and document temporal intervals, and their generalized distances

δ^* is a function between sets of intervals. We define this set-based distance by aggregating the inter-distances between each pair of elements in the two sets, assuming that a ground distance function $\delta : \Delta \times \Delta \to \mathbb{R}$ between pairs of time intervals is provided. Given δ, we define δ^* as the minimum distance between each pair of temporal intervals:

$$\delta^*(T_Q, T_D) = \min_{[s,t]\in T_Q, [u,v]\in T_D} \delta([s,t],[u,v]) \tag{2}$$

an idea borrowed from hierarchical clustering, and known as *single-link* [18]. Clearly, if δ is a generalized metric, so is δ^*, as the metric properties are preserved by the min function. With this definition, all the burden is placed on the definition of δ. Following our previous discussion, we propose the following three metrics. In all cases, as in (2), we assume that the first interval is a query interval, and the second interval is a document interval.

Manhattan Distance. Recall from Definition 2 that the timeline is a discrete set of integers. One metric that is well-suited for discrete spaces is the Manhattan distance [19] (also known as Taxicab distance, or L_1 distance). Intuitively, the Manhattan distance sums up the distances between the starting and ending times of the two intervals, resulting to zero only when the two intervals are exactly the same. We call this function δ_{sym}, to stress the fact that it is the only symmetric function (and proper metric) we consider:

$$\delta_{sym}([a,b],[c,d]) = |a-c| + |b-d|$$

Figure 2 shows five different possible cases. With δ_{sym}, knowing if an interval is from a query or a document makes no difference, since it is symmetric.

Query-Biased Hemidistance. Recall the example from Fig. 1. A user might consider the broader document more relevant because it covers the query scope. With the query-biased hemidistance, we assign a distance zero to all documents that completely cover the query scope, and a positive distance to documents that do not cover part of the query scope. Furthermore, if the query and the document scopes do not intersect, the gap between the two intervals is also added to their

distance. We call this function $\delta_{cov(Q)}$, to stress the fact that a good document covers the query:

$$\delta_{cov(Q)}([a,b]_Q,[c,d]_D) = (b-a) - (min\{b,d\} - max\{a,c\})$$

Notice from Fig. 2 how this function is not symmetric, as it is biased in favor of the first interval, i.e. the query interval.

Document-Biased Hemidistance. Symmetrically, a user might consider the broader document less relevant because it is too generic. She might consider relevant a document which falls inside the query interval or, in other words, which is covered by the query interval. The document-biased hemidistance is the opposite case as the query-biased hemidistance, hence we call this function $\delta_{cov(D)}$:

$$\delta_{cov(D)}([a,b]_Q,[c,d]_D) = (d-c) - (min\{b,d\} - max\{a,c\})$$

Again, Fig. 2 shows that this function is not symmetric, as it is biased in favor of the second interval, i.e. the document interval. It is also interesting to notice that the Manhattan distance is the sum of the two hemidistances, for any given pair of temporal intervals. Depending on the user task, any of these three metrics can be more appropriate than the others.

7 Combining the Rankings

Textual and temporal similarities cannot model the relevance in isolation better than a combination of both. In general, the textual similarity taken in isolation might be more effective than the temporal similarity taken in isolation, which implies that the two measures should be combined with different weights. The method we use is straightforward. We linearly combine the two similarity measures for each document. The resulting combined scores are, in turn, the final ranking. Similar ideas have been proposed in [13,7].

Given a query Q, we compute $sim_{kw}(Q, D_i)$ and $sim_{\delta^*}(Q, D_i)$ for each document D_i, where sim_{kw} is the keyword-only similarity. All scores are in $[0, 1]$ (normalized, if necessary) and higher for greater similarity (i.e., lower distance). This process implies transforming the results of δ^*, which are distances, to values indicating similarity. One way for doing this is with an exponential decay function (similarity decreases exponentially with distance), $sim_{\delta^*}(Q, D_i) = e^{-\delta^*(T_Q, T_{D_i})}$, which gives, by definition, scores in $(0, 1]$. If $T_Q = \emptyset$ or $T_{D_i} = \emptyset$ we set $sim_{\delta^*}(Q, D_i) = 0$. Modeling the similarity by exponential decay functions has also been studied in psychology [20]. We then compute all combined scores:

$$sim(Q, D_i) = (1-\alpha)sim_{kw}(Q, D_i) + (\alpha)sim_{\delta^*}(Q, D_i) \tag{3}$$

for a linear combination parameter $\alpha \in [0, 1]$. The final ranking is simply given by ordering the resulting set of scores. Setting α to 0 reduces the model to the keyword-only case. Setting α to 1 results in a temporal-only ranking.

8 Experimental Analysis

We evaluated the effectiveness of our proposed method using the TREC Novelty 2004 test collection,[2] consisting of 1808 documents extracted from the AQUAINT corpus, and 50 topics. The documents are news articles from three newswires (New York Times News Service, AP and Xinhua News Service), spanning a period of time of 5 years (from January 1996 through September 2000).

Queries. Topics, numbered N51-N100, comprise a *title*, a *description* and a *narrative* each. While the titles are short and concise descriptions of the query need, the descriptions and narratives are longer and truly natural language texts.

Relevance Assessments. Relevance assessments are given at the finer granularity of sentences. We abstracted from that level by simply considering relevant a document with at least one relevant sentence, obtaining in average 24 relevant documents per query. The "new sentence" assessments introduced in the Novelty track have been ignored in our experiments.

Temporal Features. We extracted temporal expressions (aka timexes) from both documents and queries with state-of-the-art NLP tools. In particular, HeidelTime [10] was used for identification, and TIMEN [11] for normalization. The first tool produced TimeML [21] documents in which timexes were annotated with TIMEX3 tags. The latter step required providing a "dct", i.e. the document (or query) creation time, to solve relative expressions (e.g. "last year"). The collection had dct's for documents but not for queries, hence we used 2013-01-01 for all queries. A set of 13 rules were used to map the normalized strings produced by TIMEN into temporal intervals, strictly following TimeML semantics. They were simple regular expressions capturing references to centuries, decades, years, months, weeks, days, as well as references to past n years, months, weeks, days, and generic past, future and current time references.

Table 1 shows statistics about timex extraction and interpretation. Temporal expressions were less frequent in queries than in documents, as we expected. However, using the descriptions and narratives from the topics gave us enough data to run our tests, resulting in 11 temporal queries.

Table 1. Temporal features in the Novelty collection

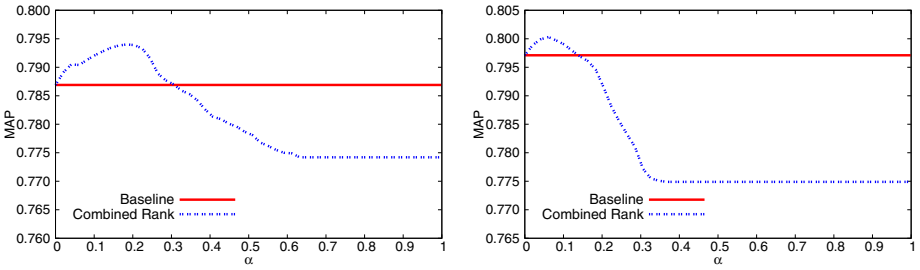
	TREC Novelty 2004 Collection		
	Documents	Topic Desc.	Topic Narr.
Number	1808	50	50
Percentage containing timexes	75%	22%	10%
Total number of timexes found	10620	14	6
Percentage of timexes mapped to intervals	81%	100%	100%

[2] `http://trec.nist.gov/data/t13_novelty.html`

8.1 Effectiveness of the Combined Ranking

In the experiment, we compared the effectiveness of combining temporal and non-temporal scores against a non-temporal ranking baseline, and we assessed the impact on temporal queries versus non-temporal queries. We selected $sim_{\delta^*_{cov(D)}}$ to model the temporal scope similarity, since it gave us the best results in terms of *mean average precision* (MAP). Lucene's default similarity[3] (based on the vector space model [17]) was used as the non-temporal, keyword-only baseline.

Sensitivity Varying α. In this test, we compared the MAP of the text-only ranking and the combined ranking, for 50 different combination parameters $\alpha \in [0, 1]$. Results are shown in Fig. 3, computed over the entire Novelty collection (all documents and all queries). Figures 3a and 3b show results when topic titles and topic descriptions, respectively, were used as textual queries. Using narratives as textual queries resulted in the worse keyword rankings, and their results are thus omitted. In all cases, the union of all the temporal expressions extracted from the topic descriptions and narratives were used as temporal queries. From the figures, it is clear how small α's improved the overall effectiveness in all cases.



(a) Text-only queries: topic titles. (b) Text-only queries: topic descriptions.

Fig. 3. Sensitivity of MAP for different combination parameters

Impact on Temporal Queries. We also compared the impact on the 11 temporal queries (i.e., such that $T_Q \neq \emptyset$) versus the entire set of queries.[4] Results are shown in Table 2. In the table, we reported *precision-at-k* (P@k), *recall-at-k* (R@k) and *MAP-at-k* (MAP@k) at different cutoff levels, when topic descriptions are used as textual queries and $\alpha = 0.06$ (the best combination weight from the previous experiment). Several conclusions can be drawn from this table. First of all, it confirms that combining textual with temporal scores improves the baseline in most cases (all the values in bold), both in terms of precision and recall. More importantly, the results obtained on the 11 temporal queries were generally higher than those obtained considering all 50 queries, even when the

[3] `https://lucene.apache.org/`

[4] Considering non-temporal queries alone would not change the rankings since we would get null temporal scores.

baseline was less effective in terms of precision (see values with $^\bullet$), or better in terms of recall and MAP (see values with *). This means that the temporal scope similarity we introduced in this paper has a higher impact on temporal queries, which, in turns, confirms the soundness of our model.

Table 2. Impact of temporal queries. Better than baseline: **bold**; better on temporal queries: *; worse on temporal queries: $^\bullet$.

Effectiveness over all 50 queries						
	Baseline			*Combined Rank*		
k	P@k	R@k	MAP@k	P@k	R@k	MAP@k
5	0.84	0.17	0.16	0.84	0.17	0.16
10	0.80	0.33	0.30	**0.81**	0.33	**0.31**
20	0.77	0.64	0.57	**0.78**	**0.65**	**0.58**
	$\alpha = 0$			$\alpha = 0.06$		

Effectiveness over 11 temporal queries						
	Baseline			*Combined Rank*		
k	P@k	R@k	MAP@k	P@k	R@k	MAP@k
5	0.83$^\bullet$	0.18*	0.17*	0.81$^\bullet$	0.18*	0.17*
10	0.79$^\bullet$	0.34*	0.31*	**0.81**	**0.35***	**0.32***
20	0.76$^\bullet$	0.66*	0.57	**0.79***	**0.69***	**0.60***
	$\alpha = 0$			$\alpha = 0.06$		

Significance Analysis. Since all score improvements were relatively small, we also performed significance analysis to strengthen our results. We run the Bootstrap Paired Test, as described in [22], using 10,000 bootstrap samples, on the 50 systems from Fig. 3b. The smallest p-value obtained was 0.05, corresponding to the system having $\alpha = 0.06$ (the best-performing one). Re-running the test by only including scores that were better than the baseline resulted in the lowest p-value of 0.04, again for $\alpha = 0.06$. This shows that there is a very low chance that the improvements given by our model are only due to chance.

9 Conclusion and Future Work

In this paper, we have studied temporal aspects of documents and queries, and we have introduced a temporal ranking model based on generalized metrics among the temporal scopes of documents and queries. We have shown that temporal scope similarities lead to effectiveness improvements only when combined with non-temporal similarity measures. This implies a *multi-faceted relevance*, in which time plays an important role. Future work will investigate ways to incorporate other dimensions, such as space, in the ranking model. We will also address the problem of efficiency: we aim at studying properties of the model that can be exploited to allow fast search in the temporal dimension and fast ranking for temporal queries.

Acknowledgments. This work has been supported in part by the Italian Ministry of Education, Universities and Research FIRB project RBFR107725 and OPLON within Smart Cities and Communities and Social Innovation project SCN_00176.

References

1. Alonso, O., Strötgen, J., Baeza-Yates, R., Gertz, M.: Temporal information retrieval: Challenges and opportunities. In: 1st Temporal Web Analytics Workshop at WWW, pp. 1–8 (2011)
2. Jones, R., Diaz, F.: Temporal profiles of queries. ACM Transactions on Information Systems (TOIS) 25(3) (2007)
3. Campos, R., Dias, G., Jorge, A.M., Nunes, C.: Enriching temporal query understanding through date identification: how to tag implicit temporal queries? In: Proceedings of the 2nd Temporal Web Analytics Workshop, pp. 41–48. ACM (2012)
4. Berberich, K., Bedathur, S., Alonso, O., Weikum, G.: A language modeling approach for temporal information needs. In: Advances in Information Retrieval, pp. 13–25 (2010)
5. Nunes, S., Ribeiro, C., David, G.: Use of temporal expressions in web search. In: Advances in Information Retrieval, pp. 580–584 (2008)
6. Snodgrass, R.T.: Temporal databases. IEEE Computer 19, 35–42 (1986)
7. Li, X., Croft, W.: Time-based language models. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management, pp. 469–475. ACM (2003)
8. Alonso, O., Gertz, M., Baeza-Yates, R.: On the value of temporal information in information retrieval. In: ACM SIGIR Forum, vol. 41, pp. 35–41. ACM (2007)
9. Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J., Pustejovsky, J.: The tempeval challenge: identifying temporal relations in text. Language Resources and Evaluation 43(2), 161–179 (2009)
10. Strötgen, J., Gertz, M.: Heideltime: High quality rule-based extraction and normalization of temporal expressions. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 321–324. Association for Computational Linguistics, Uppsala (2010)
11. Llorens, H., Derczynski, L., Gaizauskas, R., Saquete, E.: Timen: An open temporal expression normalisation resource. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (2012)
12. Metzler, D., Jones, R., Peng, F., Zhang, R.: Improving search relevance for implicitly temporal queries. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 700–701. ACM (2009)
13. Kanhabua, N., Nørvåg, K.: Determining time of queries for re-ranking search results. In: Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) ECDL 2010. LNCS, vol. 6273, pp. 261–272. Springer, Heidelberg (2010)
14. Gey, F., Larson, R., Kando, N., Machado, J., Sakai, T.: Ntcir-geotime overview: Evaluating geographic and temporal search. In: NTCIR, vol. 10, pp. 147–153 (2010)
15. Diaz, F., Dumais, S., Efron, M., Radinsky, K., de Rijke, M., Shokouhi, M.: Sigir 2013 workshop on time aware information access (# taia2013). In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1137–1137. ACM (2013)
16. Nunes, S.: Exploring temporal evidence in web information retrieval. In: Future Directions in Information Access (FDIA) (2007)
17. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)
18. Sibson, R.: Slink: an optimally efficient algorithm for the single-link cluster method. The Computer Journal 16(1), 30–34 (1973)

19. Black, P.E.: Manhattan distance. Dictionary of algorithms and data structures. US National Institute of Standards and Technology (2006)
20. Shepard, R.N., et al.: Toward a universal law of generalization for psychological science. Science 237(4820), 1317–1323 (1987)
21. Pustejovsky, J., Castano, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., Katz, G., Radev, D.: TimeML: Robust specification of event and temporal expressions in text. In: Mani, I., Pustejovsky, J., Gaizauskas, R. (eds.) The Language of time: a Reader. Oxford University Press (2005)
22. Sakai, T.: Evaluating information retrieval metrics based on bootstrap hypothesis tests. Information and Media Technologies 2(4), 1062–1079 (2007)