Temporal Query Classification at Different Granularities

Dhruv Gupta[1,2(✉)] and Klaus Berberich[1]

[1] Max Planck Institute for Informatics, Saarbrücken, Germany
[2] Saarbrücken Graduate School of Computer Science, Saarbrücken, Germany
{dhgupta,kberberi}@mpi-inf.mpg.de

Abstract. In this work, we consider the problem of classifying time-sensitive queries at different temporal granularities (day, month, and year). Our approach involves performing Bayesian analysis on time intervals of interest obtained from pseudo-relevant documents. Based on the Bayesian analysis we derive several effective features which are used to train a supervised machine learning algorithm for classification. We evaluate our method on a large temporal query workload to show that we can determine the temporal class of a query with high precision.

1 Introduction

Information needs conveyed in a time-sensitive query can only be served properly if the temporal class associated with it can be determined. Determining the temporal class of a query is an important stepping stone to larger components in a time-sensitive information retrieval system. For instance, selection of an appropriate retrieval model or deciding whether to diversify documents along time. Existing work in this direction has only relied on publication dates while ignoring temporal expressions in document contents. Temporal expressions allow us to analyze events in web collections which may not have reliable publication dates associated with them. This alleviates the problem of being restricted to the time period covered by the publication dates of the document collection. Analyzing the temporal class based on temporal expressions is challenging as (i) they are highly uncertain (e.g. `early 1990's`, `during last century`) and (ii) are present at multiple granularities (e.g., day, month, and year).

Determining the temporal class of a query has been studied before in approaches given in [2,5,6]. The approaches proposed in [2,6] however have three major problems. First, all approaches only use publication dates for a given a timestamped document collection. This may serve the purpose well for time-sensitive queries concerning only current events covered in the news. But it may be inadequate for queries covering historic events. Second, prior approaches ignore the fact that events described in a query may be periodic (e.g., `summer olympics` or `nobel prize physics`) or they may be aperiodic (e.g., `economic depression`). Third, temporal ambiguity is considered only at a single level of granularity. However temporal ambiguity may vary according to

C. Iliopoulos et al. (Eds.): SPIRE 2015, LNCS 9309, pp. 156–164, 2015.
DOI: 10.1007/978-3-319-23826-5_16

granularity. Consider, as a concrete example, the query `summer olympics tokyo athletics`. Relying only on publication dates this query would be incorrectly classified as temporally unambiguous; whereas it is temporally ambiguous at day granularity. Such an information need would be best served if these shortcomings can be overcome.

Hypothesis. By addressing the aforementioned problems we hypothesize we can improve upon the classification of time-sensitive queries containing: (i) historical events & entities; (ii) periodic events; and (iii) temporal ambiguity at a particular granularity.

We build on our earlier work [4] which suggests interesting time intervals using temporal expressions. For classifying queries we identify multiple features from Bayesian analysis of the time intervals of interest. We show the effectiveness of our proposed approach over prior work on a large testbed of time-sensitive queries.

Contributions made in this work are: (i) temporal class taxonomy taking into account multiple granularities and (a)periodicity of events (Section 4); (ii) determining time intervals as intents for temporally ambiguous queries (Section 5); (iii) effective features that outperform prior approaches (Section 6); and (iv) a large test bed of time-sensitive queries collected from previously available resources such as TREC time-sensitive queries [2], NTCIR Geo-Time queries [3] and other resources available on the Web (Section 7); which is made publicly available for future research.

2 Related Work

In this section, we describe the prior work in our context. Our work largely tries to overcome the shortcomings of work presented in [6]. The work by Jones and Diaz [6] describes a taxonomy of temporal classes for time-sensitive queries. They discuss various features derived from the distribution of document publication dates. Examples of these features are temporal clarity, kurtosis, and auto-correlation. We extend their taxonomy in our work to accomodate temporal ambiguity at different granularities, as well as (a)periodicity of events.

More recent efforts in the direction of temporal query classification have been described in works by Joho et al. [5] and Kanhabua et al. [7]. The *Temporalia* project described by Joho et al. [5] considers temporal query classification with a novel temporal taxonomy. The temporal classes they target are qualitatively labeled as *past, recency* and *future*. This has two major caveats. First, the qualitative classes leave room for ambiguity in temporal intents. For example, for `nba playoffs last week` the temporal class can either be *past* or *recent*. Second, quantitatively no information can be discerned about the *exact* time intervals the temporal class refers to. Both these problems are addressed in our work.

Detecting seasonality and periodicity associated with web-queries has also been explored by Kanhabua et al. [7]. They propose to use features acquired from web-query logs. Additionally, akin to existing approaches, they rely on features derived from signal processing on time series of publication dates from an

external document collection. These may not be adequate to detect the temporal class at different granularities, as shown in our experiments.

3 Preliminaries

We now introduce the notation used throughout the paper and the approach for identifying time intervals of interest.

Notation. Consider a document collection D. Each document $d \in D$ consists of a bag of keywords d_{text} and a bag of temporal expressions d_{time}. We let $|d_{\text{text}}|$ and $|d_{\text{time}}|$ denote the cardinalities of these bags. A temporal expression is a four-tuple, $T = \langle b_l, b_u, e_l, e_u \rangle$. Each component of T is drawn from a time domain $\mathcal{T}$ (usually $\mathbb{N}$). A temporal expression T may refer to any time interval $[b, e] \in \mathcal{T} \times \mathcal{T}$ with $b_l \leq b \leq b_u$, $e_l \leq e \leq e_u$, and $b \leq e$. We treat temporal expressions as a set of time intervals and let $|T|$ denote the number of time intervals that T may refer to.

Time Intervals of Interest to the given keyword query q are identified using the approach proposed in [4]. In a nutshell, with R as the set of pseudo-relevant documents, the approach assigns the probability:

$$P(\,[b,e] \mid q\,) = \sum_{d \in R} P(\,[b,e] \mid d\,) P(\,d \mid q\,),$$

to time interval $[b, e]$. The first probability is estimated as

$$P(\,[b,e] \mid d_{\text{time}}\,) = \frac{1}{|d_{\text{time}}|} \sum_{T \in d_{\text{time}}} \frac{\mathbb{1}([b,e] \in T)}{|T|},$$

following [1]. The second probability is estimated from the query likelihoods $P(q|d)$ under a unigram language model with Dirichlet smoothing, that is:

$$P(\,d \mid q\,) = \frac{P(\,q \mid d\,)}{\sum_{d' \in R} P(\,q \mid d'\,)}\,.$$

4 Temporal Class Taxonomy

We propose a new taxonomy taking into account additional classes for periodicity, aperiodicity, and multiple granularities (day, month, and year). It builds on the existing taxonomy proposed by Jones and Diaz [6]. The taxonomy, depicted in Figure 1, is arrived at by noting the observations explained in this section.

Atemporal queries as per [6] are time-invariant in nature. Thus, an atemporal query at year granularity also implies that it is atemporal at a finer level of granularity (day and month) and vice-versa.

Temporally unambiguous queries are those with a unique time interval of interest associated with them. If a given query is identified to be unambiguous at day granularity then it will also be unambiguous at any coarser granularity.

For instance, an unambiguous query at day level `concorde crash` is also unambiguous at year level. However, this does *not* imply that an unambiguous query at year level may necessarily be unambiguous at month or day level.

Temporally ambiguous queries are those which may have multiple time intervals of interest associated with them. Ambiguity associated with a query may lie at different granularities. A temporally ambiguous query at a finer granularity may be unambiguous at coarser granularity. However, we make the distinction that a query ambiguous at *any* granularity be deemed temporally ambiguous at that level of granularity. For example the query `summer olympics 2000 rowing` is temporally ambiguous at day level granularity. Another aspect that we investigate is the (a)periodicity of keyword queries. For example the query `summer olympics` should be classified as a periodic temporally ambiguous query. Recurring events, such as `tropical storms`, which may not have fixed periodicity are classified as aperiodic. In this work, we limit ourselves to (a)periodicity at year level. However, approach described next is equally applicable to (a)periodicity at month and day granularity.

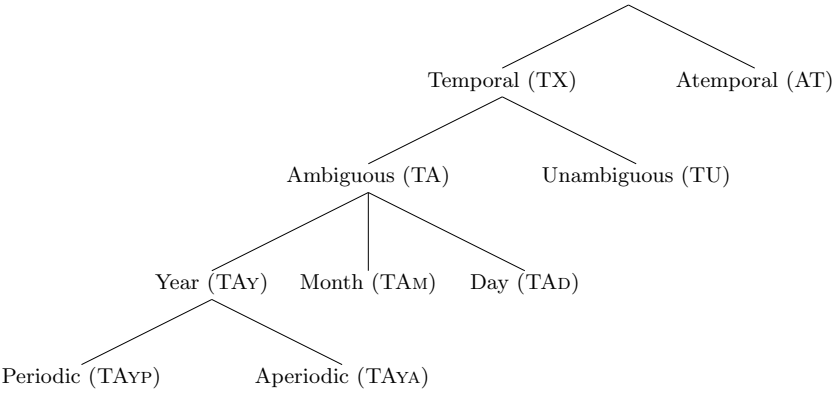


Fig. 1. Temporal class taxonomy with (a)periodicity and multiple granularity

5 Bayesian Analysis

To determine the temporal class of the keyword query q we first obtain the probability distribution of time intervals of interest at all three temporal granularities $P([b,e]|q)$. We consider time intervals of size equal to the granularity under consideration (e.g., for year granularity $[b,e]$ spans one year). We smooth $P([b,e]|q)$ with time intervals from the entire document collection D, in order to avoid the zero-probability problem:

$$\hat{P}([b,e]|q) = \lambda \cdot P([b,e]|q) + (1-\lambda) \cdot P([b,e]|D),$$

where,

$$P([b,e]|D) = \frac{1}{|D_{\text{time}}|} \sum_{T \in D_{\text{time}}} \frac{\mathbb{1}([b,e] \in T)}{|T|}.$$

Detecting Multiple Modes. The distribution $\hat{P}([b,e]|q)$ is next analyzed for multi-modality. For this we utilize a *Bayesian Mixture Model* fitted using *reversible jump Markov chain Monte Carlo* (MCMC) procedure outlined by Xu et al. [11]. The approach fits an unknown probability distribution by approximating it as mixture of Gaussian distributions. Utilizing this approach has the advantage of performing both model selection and model fitting at the same time. That is, we do not need to know the number of components in the mixture apriori. The mixture model is described as follows:

$$\hat{P}([b,e]|q) = \sum_{i=1}^{k} w_i \cdot \mathcal{N}(\mu_i, \sigma_i),$$

such that $\sum_i^k w_i = 1$; μ_i and σ_i characterize the mean and standard deviation of the normal distribution $\mathcal{N}(\mu_i, \sigma_i)$. To assess confidence of our hypothesis whether $\hat{P}([b,e]|q)$ is multi-modal, we take Bayes factor as an objective. Bayes factor is the ratio of the posterior to prior odds. If the Bayes factor exceeds 100, we consider the hypothesis, that the probability distribution under observation has multiple modes, correct.

The time intervals with the means μ_i of the components of the mixture model are the temporal categories ($S^i_{[b,e]}$) of q:

$$S = \langle S^1_{[b,e]}, S^2_{[b,e]}, \ldots, S^k_{[b,e]} \rangle.$$

6 Feature Design

After having determined the number of modes and the temporal categories from the probability distribution $\hat{P}([b,e]|q)$, we need to identify the temporal class of the keyword query. This is done by deriving features from the mixture model. The features encoded are: (i) modality, (ii) fuzzy feature, and (iii) p-value of randomness test. Next, we discuss the motivation behind the features.

Modality feature describes the number of modes identified by the Bayesian Mixture Model. The intuition is if $\hat{P}([b,e]|q)$ is unimodal ($|S| = 1$), then the temporal class should be temporally unambiguous. If the probability distribution $\hat{P}([b,e]|q)$ is multi-modal ($|S| > 1$), then it should be temporally ambiguous.

Fuzzy Feature. To analyze the temporally ambiguous query for periodicity we use the concept of fuzzy numbers. Fuzzy logic is used here to account for outlier cases in periodic events e.g. for `summer olympics` anomalous years would be [1936, 1936] and [1948, 1948]. Specifically, we capture the membership value of the time lags between the time intervals associated with different modes against a fuzzy number around the mean of the time lags ($\hat{\Phi}$).

We first identify the time lags between ordered set temporal categories Φ:

$$\Phi^i_{[b,e]} = \langle t | t \in S^{i+1}_{[b,e]} - S^i_{[b,e]} \rangle \quad \text{with,} \quad \hat{\Phi} = \frac{\sum_{i=1}^{n} \Phi^i_{[b,e]}}{n}.$$

Difference between intervals is calculated element-wise. Next we construct a triangular fuzzy number with $\hat{\Phi}$ whose membership function is given by:

$$\mu(x) = \begin{cases} \frac{1}{1+x^2} & \text{if } x \neq \hat{\Phi} \\ 1 & \text{if } x = \hat{\Phi} \end{cases}$$

The motivation is: if $\mu(x) \neq 0 \; \forall x \in \Phi$, then issued query is a periodic query with period approximately equal to that of $\hat{\Phi}$. Otherwise if $\exists x \in \Phi$ for which $\mu(x) = 0$ then query could potentially be aperiodic.

Randomness Test. For atemporal queries we check $\hat{P}([b, e]|q)$ for randomness. For this we perform a a two-tailed runs up and down test for randomness [10] on time lags. We next note the p-value of this test as a feature. This feature thus captures if the time lags are randomly generated or not.

For a given query, we construct the feature vector at day, month, and year granularity. The feature data is then subsequently used for classification via a decision tree.

7 Experimental Evaluation

7.1 Datasets

Document Collection used was The New York Times Annotated [1] corpus. Temporal annotations for it were obtained from the authors of [1]; they used TARSQI [9]. TARSQI is able to identify both explicit and implicit temporal expressions in text.

Queries. The challenging aspect of evaluating our approach was compiling a list of queries for temporally ambiguous class at different granularities. To this end we use various previously published resources [4], TREC time-sensitive queries [2], NTCIR Geo-Time queries [3], and also manually compiled some of them from the Web. Table 1 summarizes the query workload. This dataset is publicly available with an accompanying description of how it was compiled at:

http://resources.mpi-inf.mpg.de/dhgupta/data/spire2015

Table 1. Query set sizes for our evaluation setup

Set Id				Description	Size
TX	TA	TAY	TAYP	Periodic and ambiguous at year	113
			TAYA	Aperiodic and ambiguous at year	118
		TAM		Ambiguous at month	64
		TAD		Ambiguous at day	74
		TU		Unambiguous	142
AT				Atemporal	154

[1] http://www.catalog.ldc.upenn.edu/LDC2008T19

Baseline. We use the approach proposed by Jones and Diaz [6] as a baseline. We selected the best-performing temporal features from [6] to build the baseline classifier. Temporal features considered were: first order autocorrelation, kurtosis, and features derived from a burst model. We consider these features at year level granularity for time intervals of interest. Since, we are considering time intervals of interest generated by the approach in [4]; we take into account temporal expressions and publication dates at year granularity for the baseline also.

7.2 Setup

We discus various aspects related to the experimental setup next.

Parameters. For identifying time intervals of interest we considered top-50 ($|R| = 50$) pseudo-relevant documents. The mixing parameter for smoothing the distribution was set to $\lambda = 0.70$. For modality assessment we performed reversible jump MCMC procedure with 2,200 iterations with initial 200 burn-in iterations.

Implementation. All methods for feature extraction were implemented in *R*, a statistical programming language. Procedure for reversible MCMC sampling was obtained from [11], also in *R*. The decision tree classifier based on the CART algorithm was utilized from the *R* package, *rpart* [8]. The generative model for time intervals of interest was programmed in Java.

Measures. For a classification task we report the standard measures for comparing performances – *Precision, Recall* and F_1. Statistical significance of our results is reported with the p-value calculated using McNemar's test. We also show an unweighted κ statistic for the classifiers. The κ statistic measures the agreement between the observed accuracy to the expected accuracy by chance. Higher value of κ indicates better discrimination between different classes.

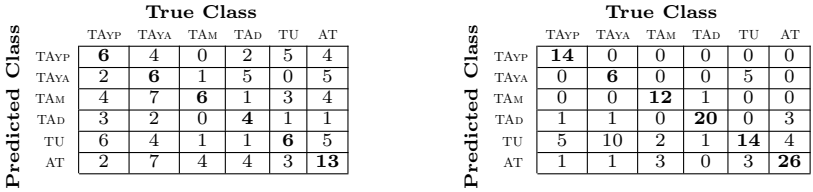
7.3 Experimental Results

Below we report the results for each temporal class. In order to accurately gauge the performance we also report the confusion matrix for our classifier. Training and test set were constructed by sampling without replacement. Train and test set split was 80% to 20% of the combined query workload (665 queries). Baseline (B) and proposed approach (A) were trained on different random samples.

Discussion. For the *temporally ambiguous* class we can classify very accurately at all levels of granularity. For the *atemporal* case we can also discern the class with high precision. However, it is relatively difficult to identify *temporally unambiguous* queries. Another class that is hard to detect is *aperiodic*. Compared to the baseline our approach performs better in all classes.

Failure Analysis. There were two classes for which our approach didn't perform well: (i) temporally unambiguous and (ii) aperiodic.

Temporally unambiguous may not have been classified precisely due to pseudo-relevance feedback. In pseudo-relevant documents it is inevitable to not



(a) Baseline (B) **(b)** Proposed approach (A)

Fig. 2. Confusion matrix for decision tree

Table 2. Statistics by class for decision trees: baseline (B) and proposed approach (A)

Statistics by Class						
	Precision		**Recall**		**F_1**	
Class	**B**	**A**	**B**	**A**	**B**	**A**
TX	0.81	**0.92**	0.79	**0.92**	0.80	**0.92**
├ TA	0.70	**0.87**	0.64	**0.71**	0.67	**0.78**
├ ├ TAY	0.45	**0.80**	0.34	**0.51**	0.39	**0.62**
├ ├ ├ TAYP	0.29	**1.00**	0.26	**0.67**	0.27	**0.80**
├ ├ └ TAYA	0.32	**0.55**	0.20	**0.33**	0.24	**0.41**
├ ├ TAM	0.24	**0.92**	0.50	**0.71**	0.32	**0.80**
├ └ TAD	0.36	**0.80**	0.22	**0.91**	0.28	**0.85**
└ TU	0.26	**0.39**	0.33	**0.64**	0.29	**0.48**
AT	0.38	**0.76**	0.41	**0.79**	0.39	**0.78**
Macroaverage	0.31	**0.74**	0.32	**0.67**	0.30	**0.69**
p-value	4.5e-2	2.2e-16				
κ-value	0.16	**0.62**				

consider other related events, which act as noise, for the keyword query in the distribution of time intervals. Some misclassified example queries are : `chernobyl soviet union` and `president nixon associated press orlando`.

Aperiodic queries were mostly misclassified as *unambiguous*. Most of the queries in the aperiodic query set comprise of famous personalities. Thus, the errors can be due to a very specific events in the corpus linked to the entity. Misclassified examples from this category are `george bush jnr`, `madrid bombing`, `muhammad ali`, and `ronald reagan`.

8 Conclusion and Future Work

We have proposed how to solve the problem of temporal query classification at multiple levels of granularity. Additionally, we can predict the periodicity of events with very high accuracy. We inspect both content temporal expressions as well as publication dates of pseudo-relevant documents given only a keyword query. Our approach considers features based on Bayesian analysis of the time intervals of interest. Experiments indicate that heuristics identified by us are able to predict the temporal class for ambiguous queries really well. In contrast, for *unambiguous* and *aperiodic* queries it is difficult to classify the class by looking

at the pseudo-relevant documents. All in all, our classifier achieves the target of temporal query classification with good accuracy.

As part of our ongoing work; we are investigating how to incorporate the temporal categories ($S^i_{[b,e]}$) of given keyword query for diversifying search results along time. As part of our future work; we plan to carry out an end to end evaluation of retrieval effectiveness when considering disambiguated temporal categories.

References

1. Berberich, K., Bedathur, S., Alonso, O., Weikum, G.: A language modeling approach for temporal information needs. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 13–25. Springer, Heidelberg (2010)
2. Dakka, W., Gravano, L., Ipeirotis, P.G.: Answering general time-sensitive queries. IEEE Trans. Knowl. Data Eng. **24**(2), 220–235 (2012)
3. Gey, F., Larson, R., Kando, N., Machado, J., Sakai, T.: Ntcir-geotime overview: Evaluating geographic and temporal search. In: Proc. NTCIR-8 Workshop Meeting, pp. 147–153 (2010)
4. Gupta, D., Berberich, K.: Identifying time intervals of interest to queries. In: CIKM 2014 (2014)
5. Joho, H., Jatowt, A., Blanco, R.: NTCIR temporalia: a test collection for temporal information access research. WWW **2014**, 845–850 (2014)
6. Jones, R., Diaz, F.: Temporal profiles of queries. ACM Trans. Inf. Syst. **25**(3) (2007)
7. Kanhabua, N., Nguyen, T.N., Nejdl, W.: Learning to detect event-related queries for web search. TempWeb 2015 at WWW 2015 (2015)
8. Therneau, T., Atkinson, B., Ripley, B.: rpart: Recursive Partitioning and Regression Trees (2014). R package version 4.1-8
9. Verhagen, M., Mani, I., Sauri, R., Littman, J., Knippen, R., Jang, S.B., Rumshisky, A., Phillips, J., Pustejovsky, J.: Automating temporal annotation with TARSQI. In: ACL 2005 (2005)
10. Wald, A., Wolfowitz, J.: On a test whether two samples are from the same population. The Annals of Mathematical Statistics **11**(2), 147–162 (1940)
11. Xu, L., Bedrick, E.J., Hanson, T., Restrepo, C.: A comparison of statistical tools for identifying modality in body mass distributions. Journal of Data Science **12**(1), 175–196 (2014)