Temporal Query Intent Disambiguation using Time-Series Data

Yue Zhao and Claudia Hauff
Web Information Systems, Delft University of Technology, The Netherlands
{y.zhao-1,c.hauff@tudelft.nl}

ABSTRACT

Understanding temporal intents behind users' queries is essential to meet users' time-related information needs. In order to classify queries according to their temporal intent (e.g. *Past* or *Future*), we explore the usage of time-series data derived from Wikipedia page views as a feature source. While existing works leverage either proprietary search engine query logs or highly processed and aggregated data (such as Google Trends) for this purpose, we investigate the utility of a freely available data source for this purpose. Our experiments on the NTCIR-12 Temporalia-2 dataset show, that Wikipedia pageview-based time-series data can significantly improve the disambiguation of temporal intents for specific types of queries, in particular those without temporal expressions present in the query string.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval
Keywords: temporal intents; disambiguation

1. INTRODUCTION

Understanding users' temporal query intents is an important step to meet their time-related information needs [8]. A query's temporal intent may be ambiguous though, as (in particular in Web search) queries often consist of 2-3 keywords only and users' information needs may be multi-faceted.

This ambiguity can be analyzed on at least two levels, the *semantic* level and the *temporal* level. On the semantic level, the same query string may refer to different concepts with very different temporal intents. For example, a query "attack the movie" issued by a user in May 2013 may either refer to (i) a 1956 film named Attack (in this case the temporal intent would be *Past*), (ii) an — with respect to 2013 — upcoming and already announced 2015 film named Attack (*Future* intent) or, (iii) war movies in general (*Atemporal* intent). On the temporal level, ambiguity is often the result of queries referring to periodically occurring events. For example, the query "NBA playoff" issued on May 1, 2013 intuitively has two strong intents: *Recency* (the 2013 playoffs were running at the time and a user may be interested in this specific playoff instance) and *Atemporal* (a user may be interested in the concept of NBA playoffs in general).

Large-scale query logs offer a rich source of temporal signals that may be useful to determine temporal intent [12, 8, 11, 7]. They are however proprietary. Although search engines release highly-processed and aggregated temporal query signals to the public (e.g. Google Trends[1]) which are often employed in research [1], we believe that a *detailed* investigation into the impact of time-series features on temporal query intent disambiguation is only possible with an openly accessible and large-scale data log. To this end, in this paper we employ the Wikipedia page view logs (containing information on how often an article has been viewed during a time period) as a source of time-series based features. We propose a two-step temporal disambiguation approach which (1) extracts a set of concepts from a query string and expands this set with related concepts, and (2) derives time-series features of all concepts found in the previous step (based on the page views of that concept on Wikipedia) which are then fed into a machine learning framework. We explore the following research question:

RQ: Do features derived from an accessible time-series data source improve the disambiguation of users' temporal intents?

2. RELATED WORK

Jones and Diaz [6] introduced the use of time-series data for the analysis of query ambiguity, exploring three types of temporal ambiguity: atemporal, unambiguous and ambiguous. Their experiments centered around a news corpus (with all documents being timestamped according to their creation date) and they exploited the change in term frequencies over time to generate time-series data for queries. Similarly, Radinsky et al. [10] leveraged time-series data generated from the New York Times collection to measure the relatedness of text. While useful in some contexts, time-series data generated from document collections may not be suitable to disambiguate the temporal intents of the general Web search user.

Query logs are a more suitable resource to disambiguate Web users' queries and research in the temporality of query logs is ongoing: Kulkarni et al. [8] monitored one hundred queries over ten weeks to learn more about the dynamics of temporal queries, while Shokouhi [12] analyzed the monthly frequency of 259 queries (issued to the Bing search engine)

[1] `https://www.google.com/trends/`

SIGIR '16, July 17-21, 2016, Pisa, Italy
 ISBN 978-1-4503-4069-4/16/07 ...$15.00.
DOI: http://dx.doi.org/10.1145/2911451.2914767.

between 2006 to 2010. Although query logs would offer the most direct evidence of users' temporal intents, they are proprietary.

As a world-wide knowledge base which contains millions of concepts with well-defined explanations, Wikipedia[2] is a meaningful resource for understanding users' queries. Whiting et al. [13] use topics in Wikipedia disambiguation pages to represent ambiguous queries & various query intents and employed the time-series data from Wikipedia's page view statistics[3] to analyze their dynamics. We take inspirations from these works and employ this data source for the specific task of temporal intent disambiguation.

Lastly, it should be pointed out that most prior works that make use of time-series data to understand query intents [6, 8, 12, 13, 7] focus on the overview dynamics, while our work attempts to disambiguate the temporal intents of each query at their particular issue time.

3. APPROACH

3.1 Temporal Disambiguation

The approach we propose for temporal query intent disambiguation is depicted in Figure 1. We consider two levels: the semantic and the temporal level. On the semantic level, employing Wikipedia concepts (a concept is operationalized as a Wikipedia page) to represent the variate intents of queries is a common choice [10, 13]. On the temporal level, each Wikipedia concept is linked to the page view statistics of its corresponding Wikipedia page. Based on the time-series data of Wikipedia concepts, several temporal features can be extracted for the estimation of query temporal intents. Our model is similar to and inspired by the model proposed in [10], the main difference being that instead of relying on time-series data directly to compare entities, we extract features from the time-series data and incorporate the features into our machine learning framework. Based on features extracted from semantic level and temporal level of queries, machine learning models are trained and leveraged to predict the probabilities of temporal queries intents where queries may belong to.

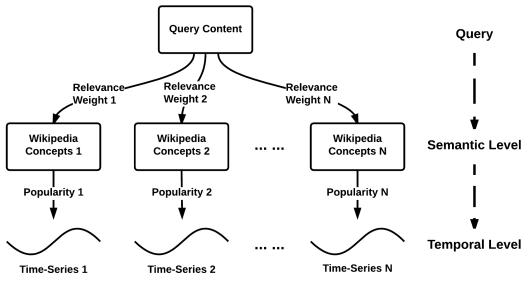

Figure 1: Conceptual overview of our approach.

3.2 Feature Extraction

We extract two types of features from the queries: (i) content features, and, (ii) time-series based features. We discuss them in turn.

[2]`https://www.wikipedia.org/`

[3]`http://dumps.wikimedia.org/other/pagecounts-raw/`

Our 209 **query-content features** (i.e. features that are not relying on any external data sources) are based on those proposed in [14]. We extract lemmas and named entities via the `Stanford CoreNLP` toolkit[4]. Since verb tenses can be good indicators of temporality, as [14] we represent the detected verbs in queries by their uppermost verb tense (*UVB_tense*) and verb tense with lemma (*tense_lemma*). For example, the query "when was television invented" has three verb features *UVB_VBD*, *VBD_be* and *VBD_invent*. Temporal expressions (TEs) are extracted with the `SUTime` module of `Stanford CoreNLP` and the relation of the detected TEs and the query issue time (assumed to be known) are encoded in five features:

- $\{ref_{past}, ref_{future}\}$: number of TEs referring to past/future times with respect to the query issue time;
- $\{same_Y, same_{YM}, same_{YMD}\}$: number of TEs referring to the same year, the same year & month, and the same year & month & day as the query issue time.

`SUTime` produces high quality temporal annotations, but is not able to detect all TEs, especially if the surrounding textual evidence is weak or misleading (e.g. in the query "When to File 2014 Taxes"[5] `SUTime` does not tag "2014" as TE; "2014" is though detected as numerical lemma by `Stanford CoreNLP`). The final three query-content features thus encode the relation between numerical lemmas[6] and query issue time:

- $\{lemY_{past}, lemY_{future}, lemY_{same}\}$: number of numerical lemmas referring to past/same/future years with respect to the query issue time.

As a concrete example, the query "NBA playoffs 2012 2013" issued May 1, 2012 will result in the following non-zero features: $\{ref_{future} = 1, same_Y = 1, lem_{future} = 1, lem_{same} = 1\}$.

The **time-series based features** and the evaluation of their impact on temporal query intent disambiguation are our main contribution in this work. We employ the `TAGME` toolkit [2] to detect Wikipedia concepts in our queries, as it has been shown to perform well for short texts. We derive 13 features per query, based on the Wikipedia concept most related to the query content[7] (i.e. having the highest relatedness score as computed by `TAGME`):

- *seasonality* [12]: represented by the cosine similarity between the time-series data and its seasonal component which is derived based on Holt-Winters additive method [3];
- *autocorrelation*: measures the periodicity of the time-series data by comparing the past 12 months of data to the same time period a year earlier;
- $\{ref_{view\_D}, ref_{view\_MD}\}$: difference between the query issue month (month/day combination) and the month (month/day combination) the concept had the most pageviews in our Wikipedia pageview traces;
- finally, the mean, standard deviation and median of the concept's time-series data are also computed.

Overall, for each query we generate 222 features based on query content and time-series data.

[4]`http://stanfordnlp.github.io/CoreNLP/`

[5]Query #199 in the Temporalia-2 dataset.

[6]To avoid noise, we only consider numerical lemmas within ±20 years of query issue time.

[7]We employ the most related concept only, as the combination of a variable number of time-series features is a non-trivial task considered in future work.

4. EXPERIMENTS

4.1 Experimental Setting

We utilize the benchmark dataset published at the NTCIR-12 Temporalia-2 task [5] for the temporal intent disambiguation (TID) subtask. It consists of *dry-run* (93 queries in total, 73 of those had their ground truth released for training purposes) and *formal-run* data (300 queries as test data). Each query has an assigned issue time. The ground truth for each query are the probabilities of the query falling into the four temporal intent classes: (*Past*, *Recency*, *Future*, *Atemporal*). For example, the query "'memorial day" (issue time May 1, 2013) has the following ground truth assignment: $(0.1, 0.0, 0.7, 0.2)$[8], i.e. the highest intent probability is assigned to the *Future* category.

In addition, for the exploratory analysis of temporal query intents, we also rely on the *formal-run* data (300 queries) released in the previous' year's (NTCIR-11) edition of the benchmark [4]. The task that year was slightly easier: queries had to be classified into a single temporal category, instead of deriving a probability distribution. The task was called temporal query intent classification ($TQIC$), a label we employ here as well to distinguish it from the TID task/data.

4.2 Data Exploration

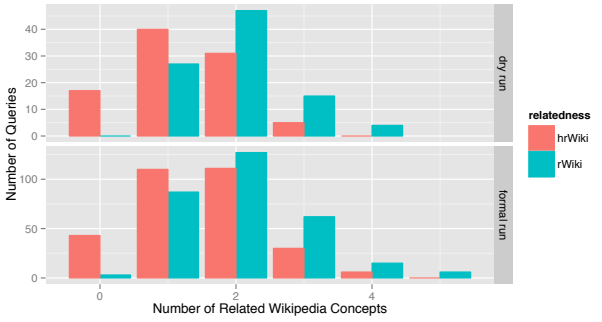

Figure 2: Overview of highly related Wikipedia concepts in the TID query set.

Concept detection: We first explored to what extent `TAGME` was able to detect concepts in our set of queries. All detected Wikipedia concepts are considered *related concepts* ($rWiKi$) while the concepts with relatedness scores ≥ 0.1 (suggested in [2]) are considered as *highly related concepts* ($hrWiki$). The number of $rWiKi$ and $hrWiki$ of queries in the dry-run dataset and the formal-run dataset of TID are shown in Figure 2. Almost every query contains at least $rWiki$ concept and more than 60% of the queries contain at least two such concepts. Highly-related concepts appear considerably fewer, about 16% of queries do not contain any $hrWiki$.

Prevalence of Temporal Expressions (TEs): To verify the intuition that queries with higher temporal ambiguity have fewer TEs, we explore the prevalence of TEs in our query sets. The results are shown in Table 1. In the older TQIC query set, more than 40% of queries contained at least one TE (which in many cases makes it easier to classify temporal intent). In the newly released TID dataset however, less than 20% of the queries contain TEs, indicating the increased need for temporal features from other sources. It is also reported that the queries with explicit temporal expressions only represent about 1.5% of all queries [9]. Therefore, it is worth to generate temporal features from other sources.

	#Queries overall	#Queries with TEs	#Queries without TEs
TQIC formal-run	300	127	173
TID dry-run	93	15	78
TID formal-run	300	57	243

Table 1: Number of Queries with TEs

Page view sparsity: We explore Wikipedia page view data as one such source. If, however, the page views of our detected Wikipedia concepts were too low, no sensible time-series data features could be generated. We define those concepts whose average (maximum) daily page views are fewer than 22 (186) as queries without sufficient page views. The cutoff values were derived from the 5^{th} percentiles of the average (maximum) daily page views across all detected concepts in our TID and TQIC datasets. In total, we find 7 queries in our datasets whose most related Wikipedia concepts (i.e. the concept to compute the time-series features from) has page views below these thresholds.

4.3 Temporal Query Intent Disambiguation

The TID taks is evaluated through: (i) the average cosine similarity between the ground truth temporal intent distribution and the predicted distribution, (ii) and the mean absolute error (MAE) which can be computed not only across all categories, but also separately for each temporal class.

Baseline: Our baseline [14] is the best performing approach for the $TQIC$ task[9], relying on query content features alone (no time-series data). As the TQIC task differs from TID (instead of predicting class labels, we now predict probabilities), we employed Ridge regression instead of Logistic regression. The parameter settings are selected by 10-fold cross-validation on the TID training data.

In order to test the effectiveness of time-series features (our main research question in this work), we extract them as described in Section 3.2. In contrast to the baseline, our model ($BrTS$) contains the additional time-series features we hypothesize will improve the temporal intent disambiguation.

We incorporate the time-series data in two ways: (i) we train a single model across all TID training queries, and (ii) we train two separate models, by splitting the TID training queries into two sets, according to whether or not they contain TEs. The results of the TID formal-run data are shown in Tables 2 and 3. The results indicate that time-series features improve the temporal disambiguation results of those queries that do not contain TEs, in particular for the categories *Future* and *Atemporal*. In contrast, the time-series features hurt the predictions for queries that do already contain TEs, especially for the *Recency* category. Comparing the single vs. two-model setup indicates that the training of two separate models aids the accuracy of the prediction. Fi-

[8]These intent distributions were derived by the benchmark organizers through crowdsourcing

[9]Official results for the TID task are not yet available.

nally, we also computed an aggregate run (*Aggr* in Table 3), which employs the baseline approach for queries with TEs and the $BrTS$ approach for those without (as this combination performs best on the training data). This combination yields a significant improvement over the baseline across all queries, not just those missing temporal expressions.

To open avenues for future work, we conducted a qualitative failure analysis on the queries where the baseline performed much worse than our $BrTS$ approach. We found that the $BrTS$ performs worse when (i) the detected Wikipedia concepts cannot capture the whole meaning of the query well, or (ii) the *Future* category has the highest intent probability and the page view log of the respective Wikipedia concepts contains multiple large irregular spikes (which lead to a larger probability of the *Past* category).

To summarize, these results suggest that 1) Time-series data can improve the disambiguation of queries with no TEs significantly, and 2) the disambiguation of users' temporal intent behind queries should be processed separately based on whether they have TEs, which is consistent with the classification of temporal queries in [7].

			Per-Class MAE			
Method	**CosSim**	**MAE**	*Past*	*Rec.*	*Fut.*	*Atemp.*
+++ All queries +++						
Baseline	0.790	0.213	0.194	0.156	0.202	0.299
BrTS	0.791	0.211‡	0.198	0.158	0.188‡	0.299
+++ Queries with TEs +++						
Baseline	0.764	0.225	0.233	0.173	0.250	0.242
BrTS	0.761	0.224	0.224	0.174	0.251	0.246
+++ Queries without TEs +++						
Baseline	0.797	0.210	0.185	0.152	0.191	0.312
BrTS	0.799	0.208‡	0.192	0.154	0.174‡	0.312

Table 2: Overview of TID results (formal-run data). A single model was trained. ‡ indicates a significant improvement ($p < 0.05$).

			Per-Class MAE			
Method	**CosSim**	**MAE**	*Past*	*Rec.*	*Fut.*	*Atemp.*
+++ All queries +++						
Baseline	0.790	0.212	0.193	0.161	0.207	0.287
Aggr	0.792	0.210‡	0.198	0.162	0.192‡	0.287
+++ Queries with TEs +++						
Baseline	0.717	0.243	0.251	0.243	0.257	0.223
BrTS	0.697	0.247	0.247	0.264	0.258	0.221
+++ Queries with no TEs +++						
Baseline	0.807	0.205	0.179	0.142	0.196	0.302
BrTS	0.809	0.202‡	0.186	0.143	0.177‡	0.302

Table 3: Overview of TID results (formal-run data). Two models were trained: one based on the training queries containing TEs, and one based on the training queries not containing TEs. ‡ indicates a significant improvement ($p < 0.05$).

5. CONCLUSIONS

In this work, we have presented our investigation into the use of time-series data, extracted from an openly accessible data source as an approximation and a proxy for large-scale query log data to predict the temporal intents of search queries.

We have provided an analysis of the NTCIR TID dataset within the context of our goal (deriving time-series data from Wikipedia page views) and found the data sparsity not to be a significant issue. Our experiments show that time-series features derived from Wikipedia page views can aid the temporal intent prediction, if sufficient care is taken to separate the easy queries (containing TAs) from the difficult ones during training.

Future work will explore additional mechanisms to incorporate the time-series features of a wider range of Wikipedia concepts (instead of the query's dominating one) and the eventual application of temporal intent prediction to the diversification and clustering of search results.

6. REFERENCES

[1] H. A. Carneiro and E. Mylonakis. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases*, 49(10):1557–1564, 2009.

[2] P. Ferragina and U. Scaiella. Tagme: on-the-fly annotation of short text fragments. In *CIKM '10*, pages 1625–1628, 2010.

[3] C. C. Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting*, 20(1):5–10, 2004.

[4] H. Joho, A. Jatowt, R. Blanco, H. Naka, and S. Yamamoto. Overview of NTCIR-11 temporal information access (temporalia) task. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies*, 2014.

[5] H. Joho, A. Jatowt, R. Blanco, H. Yu, and S. Yamamoto. Overview of NTCIR-12 temporal information access (temporalia-2) task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*, 2016.

[6] R. Jones and F. Diaz. Temporal profiles of queries. *ACM Transactions on Information Systems*, 25(3):14, 2007.

[7] N. Kanhabua, T. Ngoc Nguyen, and W. Nejdl. Learning to detect event-related queries for web search. In *WWW '15*, pages 1339–1344, 2015.

[8] A. Kulkarni, J. Teevan, K. M. Svore, and S. T. Dumais. Understanding temporal query dynamics. In *WSDM '11*, pages 167–176, 2011.

[9] S. Nunes, C. Ribeiro, and G. David. Use of temporal expressions in web search. In *Advances in Information Retrieval*, pages 580–584. Springer, 2008.

[10] K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *WWW '11*, pages 337–346, 2011.

[11] K. Radinsky, K. Svore, S. Dumais, J. Teevan, A. Bocharov, and E. Horvitz. Modeling and predicting behavioral dynamics on the web. In *WWW '12*, pages 599–608, 2012.

[12] M. Shokouhi. Detecting seasonal queries by time-series analysis. In *SIGIR '11*, pages 1171–1172, 2011.

[13] S. Whiting, J. M. Jose, and O. Alonso. Temporal dynamics of ambiguous queries. In *TAIA2015 Workshop*, volume 92, 2015.

[14] H. Yu, X. Kang, and F. Ren. Tuta1 at the ntcir-11 temporalia task. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies*, 2014.