Information Processing and Management 51 (2015) 851–868

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Generic method for detecting focus time of documents

Adam Jatowt[a,*], Ching Man Au Yeung[b], Katsumi Tanaka[a]

[a] Kyoto University, Yoshida-Honmachi, Kyoto, Japan
[b] Axon Labs Limited, Unit 308-313, Enterprise Place, Hong Kong Science Park, Shatin, Hong Kong

ARTICLE INFO

Article history:
Received 24 September 2014
Revised 14 April 2015
Accepted 5 May 2015
Available online 15 June 2015

Keywords:
Document focus time
Temporal content analysis
Temporal IR

ABSTRACT

Time is an important aspect of text documents. While some documents are atemporal, many have strong temporal characteristics and contain contents related to time. Such documents can be mapped to their corresponding time periods. In this paper, we propose estimating the focus time of documents which is defined as the time period to which document's content refers and which is considered complementary dimension to the document's creation time. We propose several estimators of focus time by utilizing statistical knowledge from external resources such as news article collections. The advantage of our approach is that document focus time can be estimated even for documents that do not contain any temporal expressions or contain only few of them. We evaluate the effectiveness of our methods on the diverse datasets of documents about historical events related to 5 countries. Our approach achieves average error of less than 21 years on collections of Wikipedia pages, extracts from history-related books and web pages, while using the total time frame of 113 years. We also demonstrate an example classification method to distinguish temporal from atemporal documents.

1. Introduction

Temporal Information Retrieval (TIR) is a subset of Information Retrieval (IR) that focuses on time-related aspects in search. TIR has been gaining recently much attention within the IR community (Alonso, Baeza-yates, Strötgen, & Gertz, 2011; Campos, Dias, Jorge, & Jatowt, 2014). The reason for this is that a relatively large fraction of search queries have temporal character. Searchers often look for information related to different temporal scopes. To properly accommodate such queries, search engines need to find documents that refer to the time periods which match time scopes underlying intents of these queries. The straightforward way is to return documents that contain dates which correspond to the temporal scope of each search query. However, such approach cannot work well in case when documents do not have any or have only few temporal expressions, neither it works in the case when the contained temporal expressions are weakly related to the core theme of documents. As an example, Fig. 1 depicts a hypothetical document that commemorates the end of World War II. It contains a mixture of sentences referring to past events and those that describe the current commemorations as well as an atemporal sentence (the last one[1]). As it can be seen, none of the sentences contains any explicit or implicit temporal expression. However, with a certain level of historical knowledge humans can position its content onto timeline as indicated on the right-hand side of Fig. 1. In fact, this cognitive process relies on using *temporal clue words* (framed by rectangles in Fig. 1) such as

* Corresponding author.
E-mail address: adam@dl.kuis.kyoto-u.ac.jp (A. Jatowt).

[1] Note that depending on particular interpretation this sentence could also be considered temporal.

http://dx.doi.org/10.1016/j.ipm.2015.05.001
0306-4573/

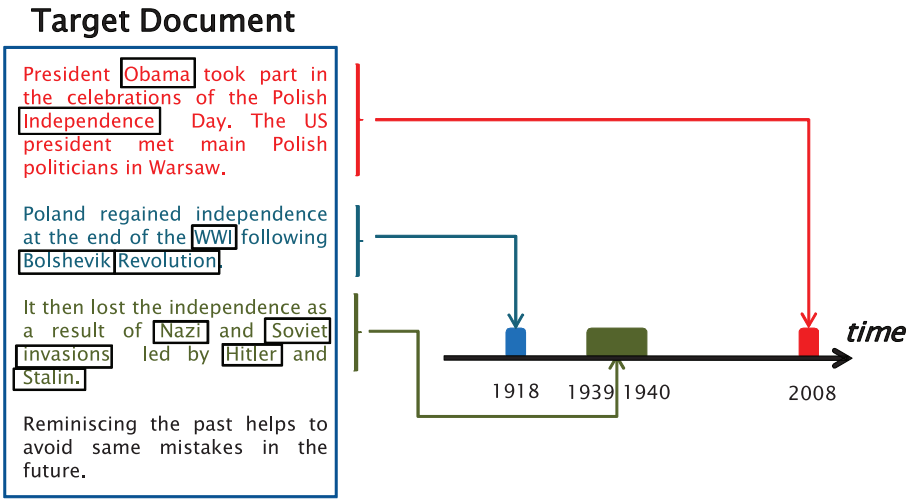

Fig. 1. Mapping content of an example document onto timeline.

"Obama," "Nazi," "Soviet," and "Stalin". The question we would like to ask in this paper is then: *Can the same process be done automatically?*

Considering that time is a key aspect of document quality, it should be beneficial to automatically categorize documents by their temporal foci and to map their content onto timeline. This would not only improve the performance of search engines in handling queries with implicit or explicit temporal intents but also help document understanding. The latter has direct applications in many text processing tasks including document summarization, information extraction, question answering and so on.

We propose estimating *document focus time*, which defines the time to which a document's content refers (as portrayed in Fig. 1). The concept of focus time is fundamentally different from the notion of *document creation time* or *timestamp* that constitutes document's basic metadata. Focus time, which means the relation of the document content to particular time periods, is essentially independent from its creation time.

We propose a range of statistical methods for automatically determining the focus time of documents by exploiting external document collections and by extracting contained direct references to time. We first compile large datasets of news articles related to different countries and then we automatically extract direct mentions of past years from their content. This allows calculating word to time associations. For example, "Nazi" and "Hitler" are strongly related to the time period from 1939 to 1945. We then propose a simple approach to relate words to years and we later extend it by considering also word's immediate contexts. In the next step, we estimate the temporal features of words to select discriminatory words useful for estimating the focus time of texts (i.e., temporal clue words like those indicated in Fig. 1). Next, we calculate the focus time of a document by aggregating the focus time of its words using various combination methods. Finally, we demonstrate how to automatically distinguish temporal documents from atemporal ones using a classification framework. We test three classifiers equipped with a range of diverse time-related features.

Note that fundamentally our approach does not require the appearance of temporal expressions in texts in order to assign documents to their corresponding time periods. That is, it can still estimate the focus time of documents that lacks explicit mentions of any dates in their content. This is an important advantage over traditional methods that rely on the existence of temporal expressions. However, since temporal expressions constitute useful signals for determining the document focus time, we also introduce a generic method that combines the purely statistical approach with the one based on processing temporal expressions. We then demonstrate that the combined method performs the best.

The remainder of this paper is organized as follows. In the next section we review the related work. Section 3 describes the methodology for calculating the document focus time. Section 4 introduces the experimental settings and Section 5 contains the results of experimental evaluation of both focus time estimation and temporal document detection. Section 6 provides a discussion of several issues related to the document focus time estimation. We conclude the paper and describe our future work in Section 7.

2. Related work

2.1. Temporal information retrieval

Temporal Information Retrieval (T-IR) (Alonso et al., 2011; Campos et al., 2014), which is a subdivision of Information Retrieval (IR), attempts to satisfy user information needs by considering not only relevance but also temporal

correspondence based on the underlying temporal factor behind search intention. As a research field, T-IR has become quite popular recently (Campos et al., 2014). It is understandable if we consider that relatively high amount of queries have temporal information needs. Nunes, Ribeiro, and David (2008) found that 1.5% queries are explicitly temporal, that is, they refer to concrete, unambiguous time points, while Metzler, Jones, Peng, and Zhang (2009) estimated that about 7% of all queries have implicit temporal intent.

Many approaches to rank documents for temporal queries (Arikan, Bedathur, & Berberich, 2009; Berberich, Bedathur, Alonso, & Weikum, 2010; Kanhabua & Nørvåg, 2009, 2010) require estimating their temporal aspects, for which they use either document metadata (timestamps) or extract explicit temporal expressions from texts. The following are the problems with the first approach. First, a document timestamp is a poor approximation of its temporal focus. This is because, obviously, documents created, for example, in 1996 do not necessarily concern events in 1996. Or, a recently document can refer to any time frame in the past or future. The document shown in Fig. 1 could be actually considered as an example of a past-focused document. While some documents such as news articles have short time lag between their creation time and focus time, this is not true for an arbitrary document. Thus, using document timestamp as a proxy of its focus time will not always work well. Second, document timestamps are not always available. For example, web documents often lack explicit timestamps or those provided cannot be trusted (e.g., "last-modified date" in web servers) (Bar-Yossef, Broder, Kumar, & Tomkins, 2004; Clausen, 2004; Oita & Senellart, 2011). The second approach that uses temporal expressions in texts also has problems and requires several assumptions. Even if a document is about a time period that corresponds to the one underlying user search intent, the temporal expressions occurring in its text can be scarce or may be missing making it impossible to correctly match the document with the user query. For example, for query "Olympics 1964," documents containing both "Olympics" and "1964" are returned. Another problem is that the appearance of a date(s) does not necessarily mean that the document is actually about the events that occurred in this date(s), since the date(s) may refer to something weakly related to the document's main theme. Following the previous example, the fact that the document contains date 1964 in its content does not mean that it is actually about the events occurring in this year.

Similar problems occur in relation to the common class of recency queries (e.g., "New York weather" or "dollar yen rate"), which are characterized by an implicit need for returning up-to-date results (Dakka, Gravano, & Ipeirotis, 2010). Again, ranking documents by their timestamps results in suboptimal performance as recently created documents may not necessarily contain information about recent events. Instead they might refer to past events or just contain obsolete content. On the contrary, documents that are actually fresh and relevant may not contain any explicit temporal expressions related to the document theme.

Close works to ours are those on identifying the temporal intent of queries (Campos, Dias, Jorge, & Nunes, 2012; Jones & Diaz, 2007; Kanhabua & Nørvåg, 2010; Metzler et al., 2009). For example, Metzler et al. (2009) proposed mining query logs to identify implicit temporal information needs by introducing a weighted measure that considers the number of times a query is pre- and post-qualified with a given year. Campos et al. (2012) demonstrated a temporal similarity measure called *GenTempEval* that associates relevant date(s) to a given query and filters out irrelevant ones based on corpus statistics rather than a document's temporal context features. There are several important differences between these works and ours. Most importantly, we work on documents instead of queries. Next, we employ a more diverse range of factors (e.g., temporal entropy/kurtosis, context-based and label propagation-based word-time associations, semantic weights, and so on) and, finally, we use news article collections as underlying knowledge bases instead of query logs or web snippets.

This paper extends our previous work (Jatowt, Au Yeung, & Tanaka, 2013). First, we propose a novel classification method for detecting temporal documents. Second, we introduce an additional word-year association approach that uses label propagation over word co-occurrence graph and a new representation of the document focus time that is based on the time intervals. Third, we report a more extensive experimentation including the usage of an additional evaluation measure, the comparison with LDA-based method and the sensitivity analysis of the proposed method with respect to the document creation date. Lastly, we offer more comprehensive discussion including the description of the Bi-Temporal Document Representation and its application for comparing and inter-relating documents.

2.2. Document timestamping and temporal information extraction

The task of document age estimation is also relevant to our work (Chambers, 2012; Garcia-Fernandez, Ligozat, Dinarelli, & Bernhard, 2011; de Jong, Rode, & Hiemstra, 2005; Kanhabua & Nørvåg, 2009; Kotsakos et al., 2014). de Jong et al. (2005) and Kanhabua and Nørvåg (2009) proposed temporal language models for document dating based on collections of time-stamped documents. For the same purpose, Garcia-Fernandez et al. (2011) used a range of features for regression functions including external knowledge information on the lifetime periods of persons mentioned in a text. Other works extended those approaches by removing the fixed time partition strategy (Kotsakos et al., 2014) or by using discriminative classifiers with features extracted from the text's time expressions (Chambers, 2012). These researches determine documents' creation dates for the purpose of estimating their age. However, as mentioned before, a document's creation date is orthogonal to the concept of document focus time since documents may refer to time periods different than their timestamps. Therefore document creation date alone cannot be enough to satisfy queries with temporal intent.

Nunes, Ribeiro, and David (2007) proposed to use the Last-Modified dates in HTTP associated with the documents linking to a given, target document in order to guess that document's age. Besides the fact that we focus on finding document focus time rather than the creation or modification time, we also use signals derived from the document content rather than ones

from the document metadata such as the Last-modified dates. In addition, we do not limit our methods to web pages but focus on any type of textual documents.

Another category of research focuses on temporal information extraction from text collections (Mazur, 2012; Strötgen & Gertz, 2010, 2012). GuTime[2] and Stanford Named Entity Recognizer[3] are examples of taggers for finding dates and other time expressions in texts. Based on temporal expression extraction, more complex systems can be built. For example, Strötgen and Gertz (2010) demonstrated a system for the extraction, querying, storage, and exploration of spatio-temporal information stored in text documents. However, as mentioned above, temporal expressions may be missing from documents or may be weakly related to the document theme.

Finally, research in automatic event extraction (Hogenboom, Frasincar, Kaymak, & de Jong, 2011) developed models that approximate linguistic phenomena by finding lexical phrases that denote certain types of actions or events. For example, the pattern <company> <buy> <company> matches phrases containing entities linked to the concept of <company> and the conjugations of verbs meaning acquisition (<buy>). "Google acquires reCaptcha" or "Skype sold to Microsoft" are examples of matched texts for such patterns. This line of research is relevant to our work since some detected event mentions could be mapped on timeline after extracting their temporal boundaries from external knowledge bases such as Wikipedia.

2.3. Topic detection and modeling

The research field of *Topic Detection and Tracking* (Allan, 2002) detects temporal patterns in document collections, for instance, by finding articles on the same events or detecting new stories. For example, Swan and Allan (2000) identified temporal features using a chi-square measure within temporal document collections to improve their browsing and summarization. Given a user query (e.g., "Bill Clinton"), Chieu and Lee (2004) extracted sentences to be placed on timelines in order to generate overviews of related events. Their method extracted dates from sentences, ranked them, and then chronologically ordered. Smith (2002) measured the co-occurrences of dates and place names in historical corpora to detect the events mentioned in document collections and analyzed how various interest measures (raw counts, chi-square, log likelihood, etc.) performed for ranking rare events.

Researches devoted to modeling topic evolution form another category of related works (Blei & Laerty, 2006; Mei, Liu, Su, & Zhai, 2006, Wang & McCallum, 2006) since they relate document content with time. Blei and Laerty (2006) and Wang and McCallum (2006) extended the Latent Dirichlet Allocation (LDA) Blei, Ng, & Jordan, 2003 to model topic evolution over time. Blei et al. assumed that topics in one year are dependent on the topics in the previous year, and Wang and McCallum assumed that each topic has its own distribution over time. Mei et al. (2006) modeled spatio-temporal patterns in weblogs. Tracking topics and their evolution in large document collections could be considered to some extent as mapping document content in time. Our work is however different as we explicitly focus on a single document rather than the collection of documents and we do not utilize document timestamps.

3. Estimating focus time

3.1. Background

First, we note that not every document has temporal character and, thus, not all documents can have their focus time estimated. A tutorial on doing matrix operations or document describing the shape of an art object have probably content that weakly connects with time. Our approach is designed for temporal documents defined as follows:

Definition 1. A **temporal document** has content that is related to time and that can be mapped to a timeline.

Note that this definition is necessarily not precise. While many documents are clearly atemporal and others temporal, for some it may be difficult to categorically state whether they are related to time. We think that the decision in such cases should depend on the particular application or user needs. In Section 5.3 we show an example classification approach for detecting temporal documents.

We define, next, the concept of the focus time of a temporal document:

Definition 2. A temporal document, d, has the **focus time** τ if its content refers to τ.

It means for example that the document describes events which occurred or persons who lived in a given time period.

In this section we describe our approach to estimate the focus time of temporal documents by using corpus-based statistics. It is composed of three steps: (1) calculating the strengths of word-time associations based on statistical knowledge derived from external document collections, (2) estimating temporal weights of words, and (3) calculating the text focus time. Step (1) defines the time periods to which individual words or, in general, any text features refer and step (2) helps identify useful discriminant features. Finally, step (3) assigns documents to their corresponding time based on the results of the previous steps.

[2] http://timeml.org/site/tarsqi/modules/gutime/ (accessed on 27/01/2015).
[3] http://www-nlp.stanford.edu/software/CRF-NER.shtml (accessed on 27/01/2015).

3.2. Measuring word-time associations

To determine word-time associations we utilize an external knowledge base. Ideally, such resource should contain many temporal references in the form of absolute dates. We believe that large datasets of news articles on diverse topics could serve as such resource. Obviously, news articles mainly describe ongoing events. However, at the same time, they also frequently refer to the past for a variety of reasons. They may provide background, explain current events by connecting them to the past ones, compare events and so on. Furthermore, news articles tend to contain many absolute temporal and spatial expression as well as often mention named entities or past events which could be easily associated with concrete dates and hence can be useful as indicators of document focus time.

Given an underlying news article collection we process it in order to form a weighted, undirected graph $G(V, E)$ containing words and their co-occurrence relationships. V denotes the set of vertices being the vocabulary of the news article collection, while E is the set of edges representing word co-occurrences. Each edge is labeled with the association weight computed by the Jaccard Coefficient (Jaccard, 1902):

$$A_{dir}(w_i,w_j) = \frac{c(w_i,w_j)}{c(w_i) + c(w_j) - c(w_i,w_j)}. \tag{1}$$

Here, $A_{dir}(w_i, w_j)$ stands for the association score between words w_i and w_j. $c(w_i, w_j)$ is the number of sentences having words w_i and w_j appearing together, while $c(w)$ is the count of sentences containing w. Note that sentence counts are used rather than document counts due to their better accuracy.

We next proceed to introduce our first method for associating words and years. If we treat dates occurring in news articles in the same way as any other words, then, a simple way to obtain the strengths of the word-year association is to directly use $A_{dir}(w, t)$ where t denotes a given time unit such as a year. Note that, for simplicity, we adopt yearly granularity and, thus, we will use interchangeably "time point" and "year" in the rest of this paper. The simple way of assigning relevance score of a year to an arbitrary word by Eq. (1) will be called *direct association* and represented as $A_{dir}(w, t)$. The direct association follows a simple hypothesis:

Word w is strongly associated with time point t if it often co-occurs with t.

The problem with the direct association is that some words may have few co-occurring years, especially, when using sentence-based co-occurrence computation. To avoid the data sparseness problem we propose to use also the term's context represented by its strongly related words in order to determine the association score between the target word and years. This resembles approaches used in NLP for computing semantic similarity between terms by utilizing their contexts and a method used for context comparison across time (Berberich, Bedathur, Sozio, & Weikum, 2009). It also bears some similarity to the approach proposed by Campos et al. (2012). We describe the intuition behind such extension as follows:

Word w has high association score with time point t if many words that frequently co-occur with w are also strongly associated with t.

In other words, we hypothesize that a term describing some event co-occurs often with other words that also describe that event. We call such extended association the *context-based association* $A_{con}(w, t)$. It is computed by the following formula:

$$A_{con}(w_i,t) = \frac{1}{|V|}\sum_{j=1}^{|V|} A_{dir}(w_j,w_i)^2 A_{dir}(w_j,t). \tag{2}$$

Note that we use square values of $A_{dir}(w_j, w_i)$ in order to decrease the impact of terms weakly associated with the target word w_i.

A further extension to the context-based approach is by performing macro-scale analysis of the association graph $G(V, E)$. For this we utilize the label propagation algorithm, which is a technique for propagating class labels in the datasets of interlinked instances based on a set of seed examples. Label propagation is performed either by an iterative Markov Chain computation or by a direct eigenvector computation (Zhu, Ghahramani, & Lafferty, 2003). We apply the latter method in which the following objective is used:

$$\min_{\boldsymbol{f}} \sum_{(i,j)\in E} l_{ij}(f_i - f_j), \tag{3}$$

where f_i is the label of vertex i and $\boldsymbol{f}$ is the vector of the labels. The objective accomplishes label propagation by forcing a pair of vertices (i, j) to have similar labels f_i and f_j if the edge weight l_{ij} is large. $\boldsymbol{f}$ is computed by taking the eigenvectors of the graph Laplacian. In our case, we treat each year as a separate class and label it with value 1, and all other words are given values 0. The algorithm then essentially "propagates" the class labels within the graph, and after reaching a convergence, it outputs the probabilities for each node as for whether the node belongs to a particular class. We use the returned probabilities as *label propagation association* between terms and years.

In the last step, we normalize each of the above listed types of association scores. The normalized values for a given word w and time point t are obtained after dividing the association scores of w with t by the geometric mean of the association scores of all other words with the time point t.

3.3. Estimating discriminative capabilities of words

We think that not all words will be equally useful to determine document focus time. For example, words that have almost uniform association with different years are clearly not useful. On the other hand, words with uneven distributions of their association scores over time should be strongly indicative of the document focus time. We then compute discriminative capabilities of words when it comes to determining document focus times. This is based on analyzing their associations over time according to the assumption:

A word has high discriminative capability for determining document focus time if it has strong association with only a few time points while having weak association with the rest of time points.

To score terms regarding their discriminative capabilities we compute *temporal entropy* and *temporal kurtosis* of terms. First, we normalize the association scores to obtain the probability distribution over time. That is, for a given word we divide its association score with a particular time point by the sum of the word's association scores with all the time points. Assuming that $P_w(t_i)$ represents the probability of a word w to be associated with time point t_i we define the temporal entropy $\omega_w{}^{temE}$ of word w as follows:

$$E_w = -\sum_i P_w(t_i)\ln P_w(t_i), \tag{4}$$

$$\varpi_w^{temE} = \max_j\left(E_j\right) - E_w. \tag{5}$$

This measure scores highly terms characterized by non-uniform probability distribution of their associations with time (compare, for example, "war" or "einstein" with "tree" or "sun"). It is similar to the one used previously in the work of Kanhabua and Nørvåg (2009) which was computed on the frequency distribution of a term over time partitions. Note that in this work we apply the temporal entropy on the distribution of word's associations with different dates as captured by the co-occurrences of the word with temporal expressions in document content.

Temporal entropy cannot however distinguish between words that have "peaks" of their distributions in near or far distance from each other, as well as, it does not consider the number of peaks. In other words, the measure treats equally temporally ambiguous and unambiguous terms. For example, words like "earthquake" or "war" may have many peaks which could be far away from one another. On the other hand, more "precise" words such as names of persons or particular events (e.g., "Einstein" or "Katrina") should be characterized by only one high peak over time. To reflect this intuition we use temporal kurtosis as a second term weighting method. It favors words that are characterized by distribution with one high peak and is defined as:

$$\varpi_w^{temK} = \frac{\sum_i (A(w,t_i) - \mu)^4}{N\sigma^4}. \tag{6}$$

In the above equation, N denotes the total number of years of the timeline we use. μ and σ respectively denote the mean and standard deviation of the word-year association distribution, $A(w, t_i)$, of a given word, w. $A(w, t_i)$ indicates any of the previously introduced word-time association measures: direct, context-based and label propagation associations (see Section 3.2). Temporal kurtosis has been also used in the context of collective memory evaluation on the view logs of Wikipedia pages devoted to events from the past (Kanhabua, Nguyen, & Niederée, 2014)

Terms with high temporal entropy or high temporal kurtosis are characterized by strong discriminative characteristics for estimating document focus time, and their appearance in a document gives a strong signal about the time periods relevant to the document. We will use temporal entropy and temporal kurtosis for term weighting and we denote either of them as $\omega_w{}^{tem}$.

3.4. Calculating document-time association

Once we obtain word-time associations and compute temporal weights of words, we can proceed to calculate the focus time of documents. The first step is to estimate the association of a target document with time. We follow an approach based on intuitive hypothesis as below:

The more words strongly associated with time point t are contained in a document d, the more it is likely that t belongs to the focus time of d.

This hypothesis is implemented by computing the weighted average over scores representing the associations of document terms with time. The weights are equal to the temporal weights introduced above. The first method we propose uses only unique words in a document:

$$S_U(d,t) = \frac{1}{|d|}\sum_{w\in d} \varpi_w^{tem} A(w,t). \tag{7}$$

$S_U(d, t)$ is the score of a time point t when using all the unique terms in the text of a document d, and $|d|$ denotes the vocabulary size of the document. The score can be regarded as the "goodness" of t to belong to the focus time of d. $A(w, t)$ represents any of the word-time association methods (see Section 3.2).

We extend the above approach by considering the frequency of a term in a document, $N(w, d)$, as follows:

$$S_{TF}(d,t) = \frac{1}{\sum_{w \in d} N(w,d)} \sum_{w \in d} \varpi_w^{tem} N(w,d) A(w,t). \quad (8)$$

Although it is common to capture the essence of a document content by considering it most frequent terms, term frequency measure alone cannot distinguish representative keywords. Besides, it often happens that frequent words are stop words that poorly describe the document content. We then propose to use another way to assign importance scores to terms in a document. Our approach resembles that used in the TextRank algorithm (Mihalcea & Tarau, 2004). We first construct undirected, weighted graph of words from a document. To build the document graph we tokenize document content, remove the stop words, and treat each remaining unique word as a vertex. An edge is added between a pair of nodes if the words represented by these nodes co-occur in at least one sentence. The score of each node is calculated recursively as follows:

$$TR_i^{(n+1)} = (1-\varepsilon) + \varepsilon \sum_{v_j \in L(v_i)} \frac{s_{ij}}{\sum_{v_k \in L(v_j)} s_{kj}} TR_i^{(n)}. \quad (9)$$

v_i denotes the i-th vertex. s_{ij} is the strength of the connection between vertices v_i and v_j and is calculated using Eq. (1), where the term counts are computed over sentences within the same document. $L(v_i)$ is the set of vertices linked with v_i. ε is a damping factor that is set to 0.85. The algorithm runs until convergence.[4] We then normalize the obtained TextRank scores and use them as importance weights of words, $\omega_w{}^{imp}$.

We next incorporate the term importance weights into Eq. (8) producing a formula that represents the document-time association scores using both the word importance and the word temporal weights:

$$S_{TR}(d,t) = \frac{1}{\sum_{w \in d} N(w,d)} \sum_{w \in d} \varpi_w^{imp} \varpi_w^{tem} N(w,d) A(w,t). \quad (10)$$

When using either of Eqs. (7), (8) and (10) the score of a time point t depends on the consensus among the time associations of the document terms. That is, if many words in a document point to the same year, the year will in result have high score and thus have more chance to be included into the focus time of the document. Moreover, when using Eq. (10) terms that are highly discriminative of the focus time and terms that are representative for the document content play stronger role (due to using importance weights).

Note that all the document-time association methods introduced so far do not explicitly use temporal expressions in target documents. However, of course, temporal expressions such as dates constitute useful signals of document's focus time. Therefore, we also propose a generic approach to incorporate the temporal expressions. This is done by extracting dates from the document content and applying Gaussian Kernel Density Estimate (Sheather, 2004) using the extracted dates. This procedure generates Gaussian distributions centered at the extracted dates. The mixture of such distributions is then considered as a date-based document-time association. We denote by $S_{DATE}(d, t)$ the score of time point t for document d calculated in this way. We then propose the extended document-time association scoring, $S_{EXT}(d, t)$, which is computed as:

$$S_{EXT}(d,t) = S_{DATE}(d,t) + S(d,t), \quad (11)$$

where $S(d, t)$ is the obtained document-time association score based on applying any of the three previously described measures (Eqs. (7), (8) and (10)).

In the last step we smooth the document-time association plots using Gaussian Kernel Density Estimate.

3.5. Computing document focus time

After computing document-time association scores we can proceed to estimate document focus time. We provide two types of document focus time representation: a time instant or a set of time instants. This categorization is similar to the temporal data types used in databases (Zaniolo et al., 1997). We call these two representations, respectively, *instant-based focus time* and *interval-based focus time*. To compute the first one we simply chose a single time point with the highest association score as the estimated focus time and denote it as $t_{foc}{}^{ins}(d)$:

$$t_{foc}^{ins}(d) = \operatorname*{argmax}_t S(d,t). \quad (12)$$

$S(d, t)$ indicates the association score between document d and time point t based on any of the measures described in Section 3.4.

Since the instant-based focus time represents the document's focus time by a single most related time point, it may be preferred in applications with high storage and processing requirements. For example, in case of search engines, adding single dates to an inverted index should have relatively small effect on the index size, while computing the temporal similarity between a query and a document should be relatively fast. On the other hand, the instant-based focus time may not be accurate enough for certain applications.

[4] Convergence is achieved when the difference among the scores in consecutive iterations is smaller than 2E−16.

Thus, we also compute the focus time as a set of time periods to satisfy the requirements of more precise temporal calculation. For implementing the interval-based focus time representation, we need to apply threshold condition for selecting years that are going to be included into the focus time. For this we calculate the difference between the maximum score and the scores of candidate time points normalized by the mean score. The years whose association scores are less than the threshold θ from the maximum score, when adjusted to the mean score, will be then included in the set of returned time points that constitute the document focus times:

$$t_{foc}^{set}(d) = \left\{ t_i : \frac{\max_t S(d,t) - S(d,t_i)}{\frac{1}{N}\sum_j S(d,t_j)} < \theta \right\}. \quad (13)$$

After selecting all the time points that satisfy Eq. (13), we reconstruct time periods by combining the adjacent time points. The final result is then either a single time period or a set of time periods.

4. Experimental settings

In this section we describe the details of the experimentation settings including explanation of used knowledge bases, test document datasets, baselines and evaluation metrics.

4.1. Preparing knowledge base

First, we have to collect a sufficient amount of temporally grounded temporal references which will serve as knowledge base for calculating the word-time association scores. For this we have collected news articles published from 1990 to 2010 from Google News Archive[5] by issuing names of several countries as queries to the search engine. We used country names rather than arbitrary queries since most events physically take place in particular countries and, by choosing such queries, we do no impose any constraint on the type of events. To diversify the data we focused on five countries: Germany, UK,[6] France, Japan, and Israel.

For each country, we collected all the returned search results with links to the original articles. Then we downloaded the article content by following the links. For a small percentage of the news articles, we were unable to collect their full texts due to subscription restrictions and in these cases, we collected abstracts instead. Next, we removed articles written in languages other than English using categorization method that applies n-gram matching (Cavnar & Trenkle, 1994). Finally, we formed five news article collections (one for each country), which contain 535k news articles in total (Germany: 87k, UK: 149k, France 110k, Japan: 97k, and Israel: 92k). Each of the above collections will be used as a knowledge base for collecting statistical information for estimating the focus times of test documents.

The collected news articles were mainly in the form of web pages. Thus to extract useful parts from them, we processed each article and removed the HTML tags, JavaScript codes, and other non-content elements. We then extracted the core part of the news articles by identifying the largest text chunk in each article. This worked well due to the relatively simple and similar layout of most news articles and allowed us not only to recover their main content but also to remove many noisy temporal expressions such as copyright dates or dates used for labeling archival content. After removing the stop words and very rare terms, the remaining text in each dataset was used for constructing word co-occurrence graphs.

We then identified temporal expressions from the news articles in each of the collection by using regular pattern expressions. We used yearly time partitioning granularity and we intentionally skipped the relative and implicit temporal expressions since resolving them is still prone to errors (Campos et al., 2014; Mazur, 2012; Strötgen & Gertz, 2012). The time frame of the extracted temporal expressions was set to [1900, 2013]. Thus we only used dates that are covered by this time period. The above time frame is long enough to cover many important historical events that happened in the countries we selected.

4.2. Document datasets

For testing our approach we prepared 15 document datasets grouped into 3 dataset categories: Wikipedia, Web, and Book. Each such category contains 5 sub-datasets of equal size, one for each country. All the documents in the datasets were written in English. The statistics of the datasets are summarized in Table 1.

4.2.1. Wikipedia dataset group

To prepare these datasets we collected 250 articles from the English Wikipedia[7] which are devoted to major historical events related to the selected countries (50 for each country). The events described in these articles occurred within the time frame of 1900–2013. The events were of different types including major wars, battles, treaties, strikes, elections, and any other

[5] http://news.google.com/archivesearch (accessed on 27/01/2015).
[6] We used disjunctive queries "United Kingdom", "Great Britain" and "UK".
[7] http://www.wikipedia.org (accessed on 27/01/2015).

Table 1
Datasets statistics (aggregated over all countries).

Dataset group	Total number of documents	Average number of sentences	Average time span of events	Mean year of events	Average number of dates
Wikipedia	250	179	3.4 years	1958	14.5
Book	735	43	4.4 years	1982	4.5
Web	819	18.3	1.3 years	1957	2.4

key events that we found for our countries. We processed all the Wikipedia articles using the CLIPS pattern library[8] and we extracted their core content by removing boilerplates and references.

We think that the Wikipedia articles on past events constitute a good source for evaluation purposes as they contain precise metadata in the form of the start and end dates of described events. As ground truth data we thus used the information in the infoboxes of the articles. These data were collected manually to ensure their accuracy.

Compared to the other two dataset categories, the Wikipedia datasets are characterized by relatively long documents that average 179 sentences with a mean of 14.5 dates per article in their content.

4.2.2. Book dataset group

To prepare the second dataset category we used two history-related books: "Timeline of World History" (Kerr, 2011) and "Timelines of History" (Ratnikas, 2012). These are the only books we could find that describe the historical events of all the countries we selected and that were available in electronic form. They cover the key historical events occurring in each year of the last century in the form of short paragraphs ordered chronologically. Since they do not provide separate timelines for our countries (only a single timeline of all the major events in the world), we extracted sentences containing the name of any of the five selected countries or their close synonyms (e.g., "British" for "UK") from each paragraph and we recorded the years of the event described in the paragraph. We then combined the sentences related to the same country that had identical years of described events into documents.

The document size of the Book dataset group is moderate averaging 43 sentences. A relatively small number of dates appear in the content (an average of 4.5 dates per document). The datasets contain documents on more recent events (the mean year is 1982) than the Wikipedia and Web datasets (1958 and 1957, respectively).

4.2.3. Web dataset group

The last dataset category was created using popular history-focused web sites. Specifically, we collected 819 texts from the websites that provide historical timelines of the selected countries: "History Orb",[9] "History World",[10] "BBC Timelines",[11] and "Infoplease".[12] We regarded each paragraph as a separate document and assigned to it the corresponding dates taken from the paragraph's title or we manually added it in case the title did not contain any dates. The average document size of this dataset is the smallest among the three dataset categories (18.3 sentences) that contain a small number of dates (on average, 2.4 dates).

4.3. Baselines

For comparing the effectiveness of our approach we prepared 3 baselines as follows.

4.3.1. Random baseline

This is the weakest baseline that randomly estimates the focus time either as a random year or a set of random years depending on whether it is used for the instant or interval-based focus time representation. In the latter case, the number of random years to be selected for each document is fixed and equals the average number of years in the ground truth data of a particular dataset. We averaged the results over 1000 random draws for each document.

4.3.2. Date-based baseline

This baseline utilizes only the absolute dates occurring in the texts disregarding the rest of content. The dates are used for generating the mixture of Gaussian distributions as described in Section 3.4. The date-based baseline is thus the same as $S_{DATE}(d, t)$.

4.3.3. LDA baseline

To construct this baseline we apply a process that uses the news article collections described in Section 4.1. First, from the news articles we extracted all the sentences containing dates with their immediately preceding and following sentences. We

[8] http://www.clips.ua.ac.be/pages/pattern (accessed on 27/01/2015).
[9] http://www.historyorb.com (accessed on 27/01/2015).
[10] http://www.historyworld.net (accessed on 27/01/2015).
[11] http://www.bbc.co.uk/history (accessed on 27/01/2015).
[12] http://www.infoplease.com (accessed on 27/01/2015).

then ran the Latent Dirichlet Allocation LDA (Blei et al., 2003) on the dataset composed of such extracts. We used 50 and 100 topics. Let C_t denote the collection of extracts that contain the mentions of a given time point t (a particular year). We calculated the average probability of a topic z in C_t using the topic probabilities of all the documents belonging to C_t. The probability distribution of the topics inside the collection C_t is given by:

$$P(z|C_t) = \frac{1}{|C_t|} \sum_{d \in C_t} P(z|d). \tag{14}$$

We assumed here that C_t represents events that happened in t; hence, $P(z|t) = P(z|C_t)$. We then constructed time point vector v_t to represent the events that occurred in t. The weight of a word w in this vector is calculated as:

$$\sum_i P(w|z_i)P(z_i|t). \tag{15}$$

Given these settings, we estimated the association score of each year with a target document. Let v_d denote the vector representation (using term frequency scoring) of a target document d. The score of a time point t for this document is equal to the similarity between v_t and v_d, where the similarity is computed using cosine similarity:

$$S_{LDA}(d,t) = cossim(v_d, v_t). \tag{16}$$

Finally, for all the three baselines described above we calculated the instant- and interval-based representations of the focus time in the same way as the one for our proposed methods, that is, using Eqs. (12) or (13). For smoothing we used Gaussian distributions with standard deviations equal to 0.6. For the interval-based representation, for all the methods, we set the same value of parameter θ equal to 0.125, which produced the best performance.

4.4. Evaluation measures

In this section we describe ways to measure the effectiveness of the document focus time estimation. For the instant-based focus time representation, we calculate the error of estimating the focus time using the following expression:

$$e(t_{foc}) = \begin{cases} \min\{|t_b - t_{foc}|, |t_{foc} - t_e|\} & \text{if } t_{foc} \notin [t_b, t_e], \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

The time period $[t_b, t_e]$ denotes the actual focus time of an input document given by the ground truth data. As expressed by Eq. (17), the error value represents the number of years between the estimated focus time point t_{foc} and the boundary of $[t_b, t_e]$. A high error means that t_{foc} is farther from the ground truth time period, while the value equal to 0 is achieved when t_{foc} falls within this time period.

When representing the focus time as intervals (i.e., interval-based focus time representation) we need to compare two sets of time periods for their overlap for each test document. We compared the correlation of two ordered lists of scores given for each year, where one list represents the results of the focus time estimation using any of our methods, and the other one represents the ground truth time period. We created the first list as follows. Each year that falls within the estimated focus time (i.e., year selected by applying Eq. (13)) is assigned value 1, and all the remaining years have values equal to 0. We used the same approach for representing the ground truth data so that the years during which a given event actually took place have values of 1, and the remaining ones have values of 0. Then we calculated the Pearson Correlation Coefficient between the two lists. The Pearson Correlation Coefficient gives value 1 for perfectly positive correlations and −1 for perfectly negative correlations. A value of 1 would indicate that the years estimated by our methods and the ground truth years are identical.

5. Experimental results

5.1. Estimating document focus time

First, we test different combinations of the proposed approaches which use document terms. These approaches do not use the dates that appear in texts. As the number of all combinations of methods is quite high we show only those whose performance is best. Table 2 displays the results for the methods that achieve the best performance for the instant-based focus time estimation, while Table 3 shows the results of the best methods for estimating the interval-based focus times. All the methods shown in these tables are statistically different from the random baseline as measured by the t-test with $p < 0.05$ (see Table 5 for the results of the random baseline).

Looking at Table 2 we notice that the best-performing method for the Wikipedia dataset group uses the context-based association, temporal entropy as temporal weight, and TextRank as term importance weight. It averages a difference of about 18.3 years between the estimated focus time year and the ground truth time period. On the other hand, for the Book datasets, the strongest method has an average error of only 16.1 years (Table 2), while for the Web dataset group the lowest error is 20.2 years. We think that these results are quite satisfactory considering rather long time span of the timeline (113 years), the relatively short average length of the described events (3.4, 4.4, 1.3 years in the Wikipedia, Book, and Web dataset

Table 2
Average error of the best method in the instant-based focus time estimation (the lower, the better).

Dataset group	Method	Error
Wikipedia	$A_{con}(w,t)$, ${\omega_w}^{temE}$, $S_{TR}(d,t)$	18.3
Book	$A_{con}(w,t)$, ${\omega_w}^{temE}$, $S_{U}(d,t)$	16.1
Web	$A_{dir}(w,t)$, ${\omega_w}^{temK}$, $S_{TF}(d,t)$	20.2

Table 3
The Pearson Correlation Coefficient of the best method in the interval-based focus time estimation (the higher, the better).

Dataset group	Method	Pearson Correlation Coefficient
Wikipedia	$A_{con}(w,t)$, ${\omega_w}^{temK}$, $S_{TR}(d,t)$	0.29
Book	$A_{con}(w,t)$, ${\omega_w}^{temE}$, $S_{TR}(d,t)$	0.30
Web	$A_{dir}(w,t)$, ${\omega_w}^{temK}$, $S_{TF}(d,t)$	0.27

Table 4
Average error of the basic methods in the instant-based focus time estimation (the lower, the better; bold font indicates the best method).

Dataset	Doc-year association	Direct assoc.	Context assoc.	Label propagation assoc.	Term weighting	
					TextRank & temporal entropy	TextRank & temporal kurtosis
Wikipedia	Score Unique	27.62	27.98	42.49	27.65	**27.47**
	Score TF	21.26	20.90	39.40	**20.55**	21.15
Book	Score Unique	20.80	**19.19**	37.73	20.56	20.65
	Score TF	20.20	**18.13**	38.07	20.10	20.10
Web	Score Unique	25.70	25.70	43.49	24.51	**24.50**
	Score TF	22.41	**21.41**	43.14	21.80	21.80

Table 5
Average Pearson Correlation Coefficient of the different combination methods in the interval-based focus time estimation (the higher, the better; bold font indicates the best method).

Dataset	Doc-year association	Direct assoc.	Context assoc.	Label propagation assoc.	Term weighting	
					TextRank & temporal entropy	TextRank & temporal kurtosis
Wikipedia	Score Unique	0.15	0.15	0.02	**0.16**	**0.16**
	Score TF	0.24	**0.27**	0.03	**0.27**	**0.27**
Book	Score Unique	0.28	**0.29**	0.11	0.28	0.28
	Score TF	0.29	**0.30**	0.10	**0.30**	0.29
Web	Score Unique	0.15	0.15	0.03	**0.16**	**0.16**
	Score TF	0.24	0.23	0.03	**0.25**	**0.25**

groups, respectively), and the variety of event types reported in the datasets for the countries we selected. Note that the results in Tables 2 and 3 do not include the approach based on explicitly detecting and normalizing dates in texts, which will be discussed later in this section. The results are then based solely on processing content words and using statistical knowledge.

Next, we can observe that the context-based association between words and time points, $A_{con}(w,t)$, works relatively well. It is the component of the best-performing methods over the Wikipedia and Book dataset groups. Therefore we think that using the year associations of terms that often co-occur with the target terms helps to improve the estimate of the focus time. On the other hand, somewhat surprisingly, the label propagation-based word-year association is not used in any of the best-performing methods listed in Tables 2 and 3 as it performs worse than the context-based association. We will later demonstrate (Tables 4 and 5) that this method consistently performs poorly. This may be due to the "leaking" connections in the word co-occurrence graph, such as the connections through words that are unrelated to any particular time frame (words with low temporal weights). In the future, we plan to experiment with the modified version of this method by removing from the graph G all words that have temporal weights lower than a certain prefixed threshold.

The temporal weights, temporal entropy (${\omega_w}^{temE}$) and temporal kurtosis (${\omega_w}^{temK}$), appear to be useful for measuring the document focus times. The TextRank measure is also useful when the document length is large, such as in the Wikipedia

datasets which contain on average longer documents than the documents in the other two dataset groups. For the Book and Web datasets, we notice that the best methods use the term frequency for document-time association, $S_{TF}(d, t)$, or one based on unique terms, $S_U(d, t)$.

We show more detailed results in Tables 4 and 5 where we report the performance of different method combinations. While we could not evaluate each different combination of all the methods we demonstrate the key ones. As a default setting we use the direct word-year association without any weights. Then we extend this setting by adding different components. Table 4 shows the average error for the instant-based focus time representation, while Table 5 shows the Pearson Correlation Coefficient for the interval-based focus time representation.

Looking at both the tables we can notice that the document-time association which uses only unique terms tends to consistently perform worse than the association based on using term frequency. The next observation is that the word-year association based on the label propagation does not perform well, as we have already mentioned above. The context-based association, on the other hand, usually provides some improvement over the direct association method. Furthermore, adding weights such as temporal entropy, temporal kurtosis and TextRank-based semantic method often allows to boost the performance to certain extent. However, in few cases the approach without any weights works well, too (e.g., in the case of the Book datasets). The temporal entropy performs in a similar way to the temporal kurtosis.

Tables 6 and 7 compare the results of the selected combination of the proposed methods against the results generated by the baselines. For comparison we chose a combination method that uses the context-based association ($A_{con}(w, t)$), temporal kurtosis (${\omega_w}^{temK}$) as temporal weight, and TextRank scoring ($S_{TR}(d, t)$) as importance weights. We also use its extended version that incorporates dates in texts by applying Eq. (11). The proposed methods are called, respectively, *Prop* and *PropExt*.

Looking at Table 6, we observe that *Prop* outperforms the baselines for all the datasets except when the date-based baseline is applied on the Wikipedia datasets. We think that this is because the Wikipedia articles contain many dates (see Table 1); hence, the straightforward date-based baseline naturally performs better. However, as mentioned before, many texts about the past may not contain any temporal expressions or may contain only a few of them making it difficult or impossible to estimate the focus time using straightforward approaches. Hence, for the Book and Web datasets the date-based baseline performs very poorly, and in these cases, an extended approach is preferred over the one that only relies on processing temporal expressions.

Indeed, we observe that *PropExt* performs best in all the cases listed in Tables 6 and 7. For the Wikipedia dataset group (Table 6) the error averages only 2.83 years. For the other datasets the performance is also improved when compared to the *Prop* results. However, after applying the Tukey HSD test ($p < 0.05$) to the results shown in Tables 6 and 7, we found a lack of significant difference between the date-based baseline and *PropExt* applied on the Wikipedia datasets. All other comparisons between the results of the baselines and the proposed methods were found to be significantly different. This implies that when enough dates are present in a text, as in the case of the Wikipedia documents, the improvement over the date-based approach may not be significant. The benefit due to applying the extended method (*PropExt*) remains significant for the other datasets with fewer dates in their content. When looking at the performance of the LDA baseline, we conclude that it cannot reliably estimate focus times achieving, in the best case, only 0.07 value of the correlation for the interval-based focus time estimation. The random baseline performs worst since it has no correlation with the ground truth at all.

5.2. Effect of time distance

We next investigate the effect of time distance on the accuracy to determine whether there is any difference in performance depending on the focus time of the test documents. We binned documents in our datasets so that each bin contains only documents that refer to events in the same decade. Next, we calculated the average error of the *Prop* method for the documents in each bin.

Fig. 2 shows the boxplots of the average error of the instance-based focus time estimation per decade. We also display the mean error by solid blue lines. Looking at the error rates for different decades, we notice that the error decreases in time when moving toward the most recent decade. This means that our approach performs better for the more recent decades. This is mainly because in the knowledge bases we use, the recent events are referenced more frequently than distant events, such as those around the beginning of the last century. There is simply less information on the distant events in the news articles that we collected. Recall that the collected news articles were published from 1990 to 2010. The increase in the error for the latest decade (2010s) reflects the mismatch between the timestamps of the news articles used for creating the knowledge bases and the test collections. Since the news articles from which we extracted dates were published before 2010, they have few references to the events in [2010, 2013] other than future plans. When looking more carefully at Fig. 2, we realize that the key decade is 1940s. It seems to divide the “recent past” from the “remote past”. Many critical events for the countries we selected happened around that time, e.g., World War II, NATO’s formation, the Marshall Plan. Since these events are frequently cited in the knowledge base, the performance on this decade is very good across all datasets.

5.3. Detecting temporal documents

Definition 1 explained the concept of temporal documents. We demonstrate here how to build classifiers for detecting temporal documents and we show the results of experiments performed on Wikipedia articles. The objective of the

Table 6
Average error for the instant-based time focus estimation (the lower, the better; bold font indicates the best method).

Dataset group	Random	Date-based	LDA (50 topics; 100 topics)	Prop	PropExt
Wikipedia	36.5	3.02	29.4; 27.2	18.3	**2.83**
Book	39.3	48.1	40.4; 37.3	23.5	**20.4**
Web	40.5	53.4	40; 41.4	23.6	**20.7**

Table 7
Pearson Correlation Coefficient for the interval-based time focus estimation (the larger, the better; bold font indicates the best method).

Dataset group	Random	Date-based	LDA (50 topics; 100 topics)	Prop	PropExt
Wikipedia	0	0.65	0.07; 0.1	0.29	**0.66**
Book	0	0.01	0.05; 0.04	0.25	**0.30**
Web	0	0.06	0.03; 0.02	0.26	**0.41**

experiments is to show that separating temporal from atemporal documents is feasible with good accuracy using small number of features. We use the following features.

F1. The ratio of verbs in past tense to the total number of verbs.
F2. The mean date of all dates occurring in text. If the text has no date, we use "2012" by default.
F3. The variance of dates occurring in text.
F4. The mean temporal entropy of words.
F5. The variance of temporal entropy of words.
F6. The average temporal kurtosis of words.
F7. The variance of temporal kurtosis of words.
F8. The mean date of the interval-based focus time calculated by *Prop* method.
F9. The variance of dates constituting interval-based focus time as calculated by *Prop* method.
F10. The maximum score among the scores calculated by Eq. (10).
F11. The mean score of the scores calculated by Eq. (10).

The motivation behind Features F2 and F8 is that temporal documents tend to be related to events from the past. On the other hand, the rationale behind choosing Features F3 and F9 is an assumption that documents about historical event tend to have narrow time focus centered on the years related to the discussed event. Features F4, F5, F6 and F7 are used since we expect temporal documents to contain many words with relatively high temporal weights. Lastly, temporal documents should be characterized by high values of F10 and F11.

As the training and testing data we use a part of the Wikipedia datasets described in Section 4.2. In addition, we prepare also additional collections containing manually selected Wikipedia articles about concurrent entities (persons, places, etc.) for each country. The former (i.e., the Wikipedia dataset used in the previous experiments) is used now as the source of temporal documents and will be called *temporal collection*. The latter (i.e., the newly selected Wikipedia articles on the current entities) is used as source of atemporal documents and will be called *atemporal collection*. Based on these collections we construct the following dataset pairs (values in parentheses indicate the total number of documents in each class aggregated over all countries):

- **Dataset group A**. The most history-related sections of documents in the temporal collection (237) vs. the abstract of documents from the atemporal collection (435).
- **Dataset group B**. "History" sections of documents in the atemporal collection (285) vs. the abstract of documents from the atemporal collection (285).
- **Dataset group C**. The whole content of documents in the temporal collection (237) vs. the whole content of documents in the atemporal collection (435).

Note that although the dataset group B contains document parts extracted only from the atemporal collection, we use the history sections as documents in the temporal class. Many Wikipedia articles contain "History" sections. These sections can be naturally treated as temporal documents. However, the documents of the Wikipedia dataset, which we use here, have few such sections, since these datasets contain articles about past events and thus, more or less, their whole content is actually about the past (hence no need for editors to create explicit "history" section in these articles). Therefore, when creating the dataset group A, if a given document lacks an explicit history section, we select a section that was most related to the "history" of the described events or entities (e.g., "background" or "overview" sections).

In total we use 15 datasets (3 groups $*$ 5 countries). All the Wikipedia articles were processed by using the CLIPS library and subject to processing in order to extract core content and segment it into paragraphs. As classifiers we use Linear

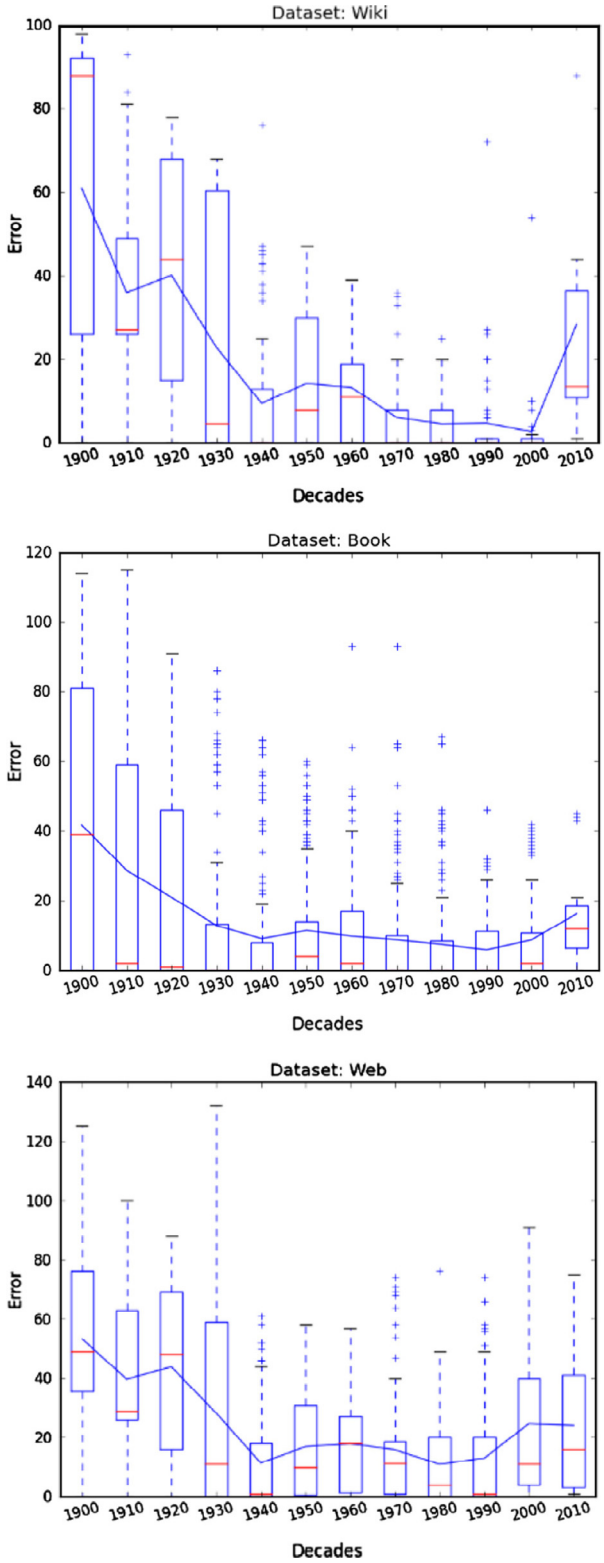

Fig. 2. Boxplots and mean error (blue line) per decade for selected method *Prop* on Wikipedia (top), Book (middle) and Web (bottom) dataset groups. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Discriminant Analysis Classifier (LDAC) (Hastie, Tibshirani, & Friedman, 2009), Perceptron and Large-scale Linear Support Vector Regression Classifier (LLSVRC) (Ho & Lin, 2012) with L2-regularized L2-loss support vector classification. Tables 8–10 show average precision, recall, F1 measure and accuracy based on the results of the 4-fold cross validation test. We notice that the classification accuracy is quite high which indicates that separating temporal from atemporal documents is feasible. Different classifiers perform more or less similarly. The best results are obtained for the dataset group C, which may be due to the longest length of documents.

After performing feature ranking using the Recursive Feature Elimination algorithm (Guyon, Weston, Barnhill, & Vapnik, 2002) we found that, in majority of the cases, the most important feature was the ratio of past verbs to the total number of verbs (Feature F1). The accuracy of classifiers (listed in the order: LDAC, Perceptron and LLSVRC) without this feature was: 0.81, 0.71, 0.81 for the dataset group A, 0.73, 0.64, 0.74 for the dataset group B and 0.88, 0.85, 0.88 for the dataset group C. F5 (variance of temporal entropy of terms) was the second most important feature followed by F3 (variance of dates in text). We also tested the combined impact of the date-related features F2 and F3. The accuracy without these features was 0.89, 0.80, 0.87 for the dataset group A, 0.86, 0.83, 0.86 for the dataset group B and 0.88, 0.73, 0.86 for the dataset group C, respectively. The results indicate that it is possible to successfully determine temporal documents from atemporal ones without using any mentions of dates in the text.

Although high-accuracy classification appears to be feasible in the view of our experiments, we want to emphasize that the reported results are generated based on only one type of documents, Wikipedia articles, and the performance can be different for other document types. In the future, we plan to analyze classification accuracy of other document types or genres.

6. Discussions

In this section we describe several related issues to the process of document time estimation.

6.1. Temporal vs. atemporal documents

Clearly, the methods proposed in this paper can be applied only to temporal documents. While we show the classification approach for detecting temporal documents, we note that the distinction between temporal and atemporal documents is as philosophical as the discussion of the nature of time itself. Most documents contain some temporal information or refer to some particular time period in the past. However, broadly speaking, we can still argue that certain documents are clearly not temporal (e.g., guidelines, rules, or directions to a particular place), while others are temporal (e.g., the descriptions of past events, biographies of famous persons, or timelines).

6.2. Data sources and "history spaces"

If we assume the existence of different "histories" such as the histories of countries, regions, concepts, and scientific areas, then the focus time of text documents could be estimated using any of such diverse historical spaces. Obviously, for an article about past events related to, for instance, France we would obtain the best results when using a news article collection that is also related to France rather than one related to, for example, Japan. On the other hand, for a document unrelated to any particular country or any physical location, such as one on past discoveries in physics, one could apply a "physics history space" using the corresponding knowledge base. Note that in the experiments we used news datasets related to particular countries. This allows us to position documents on timelines that represent the histories of the corresponding countries. An extension of this approach is the application of country-independent data sources or, perhaps, the combined sources for representing the entire world history.

6.3. Temporal modifiers and clue words

Temporal modifiers and other similar expressions provide temporal clues on the order and chronological position of events. For example, "after," "following," "before," "during," "from," and "between x and y" add temporal criteria for the content that occurs in the neighborhood of these expressions. We could model the contributions of such words by applying different time distributions (e.g., exponential, uniform, etc.) that depend on the semantics of the temporal modifiers in a similar way as in Jatowt and Au Yeung (2011).

6.4. Time granularity and events without time references

In the future we plan to utilize other time granularities than yearly granularity. Using finer granularity expressions such as months or even days could help to more precisely quantify the document focus time and make it even possible to find documents on a particular day or month. However, the problem with such choice is how to accumulate large enough amount of references to finer granularity time periods.

Table 8
Classification results for the dataset group A (aggregated over all countries).

Classifier	Precision	Recall	F_1	Accuracy
LDAC	0.86	0.87	0.86	0.88
Perceptron	0.85	0.84	0.84	0.86
LLSVRC	0.87	0.89	0.87	0.88

Table 9
Classification results for the dataset group B (aggregated over all countries).

Classifier	Precision	Recall	F_1	Accuracy
LDAC	0.87	0.87	0.87	0.87
Perceptron	0.86	0.86	0.86	0.86
LLSVRC	0.87	0.87	0.87	0.87

Table 10
Classification results for the dataset group C (aggregated over all countries).

Classifier	Precision	Recall	F_1	Accuracy
LDAC	0.90	0.90	0.90	0.91
Perceptron	0.89	0.87	0.88	0.89
LLSVRC	0.90	0.89	0.89	0.90

Another issue concerns estimating the focus time of documents about events that are typically never associated with dates. Precise starting and ending dates may not be known for certain events, or, they may be fuzzy or unreliable. To solve this problem we might have to use a special type of temporal inference.

6.5. Document focus time and creation time

As we noted before, the document focus time is orthogonal to the document creation time. We can then represent both times in a similar way as the Bi-Temporal Data Representation in database research (Zaniolo et al., 1997) where *transaction time* is analogous to the document creation time, and *valid time* is analogous to the document focus time. Such representation would place documents within the space of two orthogonal timelines assuming the documents are temporal and that both the document creation time and the focus time are known. Note that the focus time extends into the future in such representation. The creation time is obviously bounded by the moving point "now."

Fig. 3 shows example documents positioned in such a representation. Documents are represented here either as points (single time point) or as horizontal lines (set of time points). The lines are horizontal under the assumption of the instantaneous creation of the content. For example, d_3 is a more recent document than d_2, and its focus time period partially overlaps that of d_2, while d_4 is more recent than d_3 and has an instant focus time.

The comparison of the two types of time (document focus time t_{foc} and document creation time t_{cre}) allows documents to be placed in the different time categories based on their position in relation to the dotted, diagonal line. Documents are about the past if $t_{foc} < t_{cre}$ (documents placed above the diagonal line such as d_3 and d_2), about the present if $t_{foc} = t_{cre}$ (these are documents placed on the diagonal in Fig. 3, such as d_4 and d_7), and documents are about the future if $t_{foc} > t_{cre}$ (ones below the diagonal such as d_1 and d_5). Certain documents might belong to more than one category such as d_6. Documents can also be temporarily related to each other by applying relations between their focus time intervals as introduced by Allen (Allen, 1983) (e.g., d_i after d_j, d_i meets d_j or d_i during d_j).

We think that the Bi-Temporal Document Representation could complement the semantic representation of documents (e.g., vector space model) and make it possible to measure *temporal similarity* between two documents or between a query and a document in parallel to their content similarity.

6.6. Other applications

Finally, we think that the document focus time can be applied in other applications besides T-IR such as ones listed below:

- Improving temporal annotation and extraction of documents.
- Detecting references to the past in texts.
- Improving sentence ordering in document summaries.
- Image dating using focus times of surrounding texts of images.

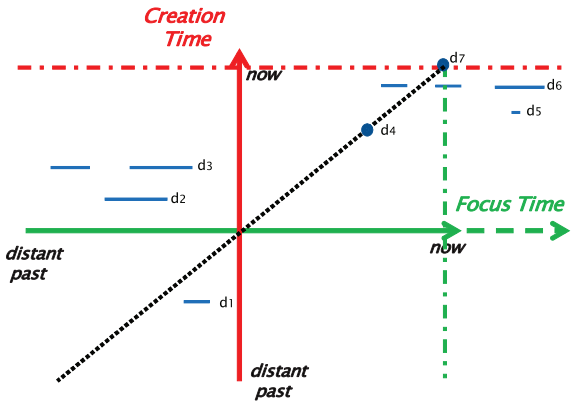

Fig. 3. The Bi-Temporal Document Representation for example documents.

- Estimating temporal analogies between distant events/entities.
- Supporting *computational history* (Au Yeung & Jatowt, 2011) and *culturomics* (Michel et al., 2011) research fields.

7. Conclusions

Time matters greatly in our lives. It is also an important aspect of texts. We think that properly estimating the content time of temporal documents should improve temporal information retrieval and strengthen our means of analyzing and understanding documents and temporal references in texts. In this paper, we describe the concept of document focus time and provide a range of methods for its estimation. Our approach harnesses corpus statistics, especially, it uses absolute references to past years in news articles. The intriguing characteristic of our proposal is that it also works for documents which do not contain any temporal expressions. Besides estimating document focus time we also demonstrate the classification approach for detecting temporal documents. The experimental evaluation indicates that the proposed methods provide satisfactory results over diverse sets of documents.

In future we will focus on the above mentioned plans and will also formally define operations in temporal search models that incorporate both document focus time and its timestamp.

Acknowledgments

This research was supported in part by MEXT Grant-in-Aid for Young Scientists B (#22700096) and by the JST research promotion program Sakigake: “Analyzing Collective Memory and Developing Methods for Knowledge Extraction from Historical Documents”.

References

Allan, J. (Ed.). (2002). *Topic detection and tracking: Event-based information organization*. USA: Kluwer Academic Publishers.
Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *CACM, 26*, 832–843.
Alonso, Omar, Baeza-yates, Ricardo, Strötgen, Jannik, & Gertz, Michael, (2011). Temporal information retrieval: Challenges and opportunities. In *TWAW 2011*.
Arikan, I., Bedathur, S. J., & Berberich, K. (2009). Time will tell: Leveraging temporal expressions in IR. In *WSDM 2009*.
Au Yeung, C.-M., & Jatowt, A. (2011). Studying how the past is remembered: Towards computational history through large scale text mining. In *CIKM 2011* (pp. 1231–1240).
Bar-Yossef, Z., Broder, A. Z., Kumar, R., & Tomkins, A. (2004). Sic transit gloria telae: Towards an understanding of the web's decay. *Proceedings of the 13th international conference on world wide web (WWW '04)* (pp. 328–337). New York, NY, USA: ACM.
Berberich, K., Bedathur, S. J., Sozio, M., & Weikum, G. (2009). Bridging the terminology gap in web archive search. In *Proceedings of WebDB 2009 workshop*.
Berberich, K., Bedathur, S. J., Alonso, O., & Weikum, G. A. (2010). Language modeling approach for temporal information needs. In *ECIR 2010* (pp. 13–25).
Blei, D. M., & Laerty, J. D. (2006). Dynamic topic models. In *ICML 2006* (pp. 113–120).
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.
Campos, R., Dias, G., Jorge, A., & Jatowt, A. (2014). Survey of temporal information retrieval and related applications. *ACM Computing Surveys, 47*(2), 1–41 (ACM Press).
Campos, R., Dias, G., Jorge, A. M., & Nunes, C. (2012). GTE: A distributional second-order co-occurrence approach to improve the identification of top relevant dates. In *CIKM 2012* (pp. 2035–2039).
Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. In *SDAIR 1994* (pp. 161–175).
Chambers, N. (2012). *Labeling documents with timestamps: Learning from their time expressions* (pp. 98–106). Association for Computational Linguistics, ACL.
Chieu, H. L., & Lee, Y. K. (2004). Query based event extraction along a timeline. In *SIGIR 2004*.
Clausen, L. R. (2004). Concerning Etags and datestamps. In *Proc. IWAW2004 workshop*.
Dakka, W., Gravano, L., & Ipeirotis, P. (2010). Answering general time-sensitive queries. In *TKDE 2010*.
Garcia-Fernandez, A., Ligozat, A.-L., Dinarelli, M., & Bernhard, D. (2011). When was it written? Automatically determining publication dates. In *SPIRE 2011* (Vol. 7024, pp. 221–236).
Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.
Ho, C.-H., & Lin, C.-J. (2012). Large-scale linear support vector regression. *Journal of Machine Learning Research, 13*, 3323–3348.
Hogenboom, F., Frasincar, F., Kaymak, U., & de Jong, F. (2011). An overview of event extraction from text. In *Workshop on detection, representation, and exploitation of events in the semantic web (DeRiVE 2011) at ISWC 2011* (pp. 48–57).
Jaccard, P. (1902). Lois de distribution florale. *Bulletin de la Socíeté Vaudoise des Sciences Naturelles, 38*, 67–130.
Jatowt, A., & Au Yeung, C.-M. (2011). Extracting collective expectations about the future from large text collections. In *CIKM 2011* (pp. 1259–1264).
Jatowt, A., Au Yeung, C. M., & Tanaka, K. (2013). Estimating document focus time. In *CIKM 2013* (pp. 2273–2278).
Jones, R., & Diaz, F. (2007). Temporal profiles of queries. *TOIS: ACM Transactions on Information Systems, 25*(3).
de Jong, F. M. G., Rode, H., & Hiemstra, D. (2005). Temporal language models for the disclosure of historical text. In *AHC'05* (pp. 161–168).
Kanhabua, N., & Nørvåg, K. (2009). Using temporal language models for document dating. In *MLKDD 2009* (pp. 738–741).
Kanhabua, N., & Nørvåg, K. (2010). Determining time of queries for re-ranking search results. In *ECDL 2010* (pp. 261–272).
Kanhabua, N., Nguyen, T. N., & Niederée, C. (2014). What triggers human remembering of events? A large-scale analysis of catalysts for collective memory in Wikipedia. In *Proceedings of JCDL 2014* (pp. 341–350).
Kerr, G. (2011). *Timeline of world history*. Canary Press.
Kotsakos, D., Lappas, T., Kotzias, D., Gunopulos, D., Kanhabua, N., & Nørvåg, K. (2014). A burstiness-aware approach for document dating. *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval (SIGIR '14)* (pp. 1003–1006). New York, NY, USA: ACM.
Mazur, P. (2012). *Broad-coverage rule-based processing of temporal expressions* (pp. 1–245). PhD thesis, Australia Macquarie University.
Mei, Qiaozhu, Liu, Chao, Su, Hang, & Zhai, ChengXiang, (2006). A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW2006* (pp. 533–542).
Metzler, D., Jones, R., Peng, F., & Zhang, R. (2009). Improving search relevance for implicitly temporal queries. In *SIGIR 2009* (pp. 700–701).
Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176–182.
Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *EMNLP* (pp. 404–411).
Nunes, S., Ribeiro, C., & David, G. (2007). Using neighbors to date web documents. *Proceedings of the 9th annual ACM international workshop on web information and data management (WIDM '07)* (pp. 129–136). New York, NY, USA: ACM.
Nunes, S., Ribeiro, C., & David, G. (2008). Use of temporal expressions in web search. *Proceedings of the ECIR'08. Lecture notes in computer science, 4956/2008* (pp. 580–584). Springer-Verlag.
Oita, M., & Senellart, P. (2011). Deriving dynamics of web pages: A survey. *TWAW (temporal workshop on web archiving), Hyderabad, India, March 2011*.
Ratnikas, A. (2012). *Timelines of history* (Kindle ed.).
Sheather, S. J. (2004). Density estimation. *Statistical Science, 19*(4), 588–597.
Smith, D. (2002). Detecting and browsing events in unstructured text. In *SIGIR 2002*.
Strötgen, J., & Gertz, M. (2010). TimeTrails: A system for exploring spatio-temporal information in documents. In *VLDB 2010* (pp. 1569–1572).
Strötgen, J., & Gertz, M. (2012). Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 1–30.
Swan, R., & Allan, J. (2000). Automatic generation of overview timelines. In *SIGIR 2000* (pp. 49–56).
Wang, X., & McCallum, A. (2006). Topics over time: A non-markov continuous-time model of topical trends. In *KDD 2006* (pp. 424–433).
Zaniolo, C., Ceri, S., Faloutsos, C., Snodgrass, R. T., Subrahmanian, V. S., Zicari, R., et al. (1997). TSQL2. *Advanced database systems*. Morgan Kaufmann.
Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML 2003* (pp. 912–919).