Estimating Document Focus Time

Adam Jatowt[1,2], Ching-Man Au Yeung[3] and Katsumi Tanaka[1]

[1]Kyoto University
Yoshida-Honmachi, Sakyo-ku
606-8501 Kyoto, Japan
{adam, tanaka}@dl.kuis.kyoto-u.ac.jp

[2]Japan Science and Technology Agency
4-1-8 Honcho, Kawaguchi-shi,
Saitama, 332-0012 Tokyo, Japan

[3]Noah's Ark Lab, Huawei
Units 525-530, Core Building 2
Hong Kong Science Park, Shatin,
Hong Kong
albert.auyeung@huawei.com

ABSTRACT

Temporality is an important characteristic of text documents. While some documents are clearly atemporal, many have temporal character and can be mapped to certain time periods. In this paper, we introduce the problem of estimating focus time of documents. Document focus time is defined as the time to which the content of a document refers to and is considered as a complementary dimension to its creation time or timestamp. We propose several estimators of focus time by utilizing external knowledge bases such as news article collections which contain explicit temporal references. We then evaluate the effectiveness of our methods on diverse datasets of documents about historical events in five countries.

Categories and Subject Descriptors

I.7.5 [**Document Capture**]: Document Analysis; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Document focus time, temporal content analysis, temporal IR

1. INTRODUCTION

Humans refer to the past for various reasons such as providing explanation for the present, motivating their actions, warning against possible dangers, emphasizing durability and trust, etc. References to the past in text can take diverse forms such as mentions of past persons, historical places or historical events. Considering the importance of history and time in our lives, it should be beneficial to provide automatic means for categorizing documents according to their temporal foci and for mapping their content onto the timeline. This would lead to better document understanding, and would also improve the performance of search engines in handling user queries with implicit or explicit temporal intent [1,2,4,5,9]. It has been actually reported that a significant number of search queries on the Web contain underlying temporal intent [7,14]. In this paper we introduce the concept of *document focus time* which defines the time to which document content refers to. Note that the concept of the focus time is fundamentally different from the notion of the document creation time which constitutes basic document's metadata. The focus time means the relation of document content to particular time periods and is essentially independent from the document creation time.

We propose a set of methods for automatically evaluating focus time of documents that make use of collection statistics and, in particular, extracted direct references to time. Our approach is as follows. We first compile large datasets of news articles related to a few selected countries. Direct mentions of past years are then automatically extracted from the news articles. This allows calculating word associations with time such as word to year relationships. For example, "Nazi" and "Hitler" would be strongly related to the time period of the World War II as frequently co-occurring with dates within the period 1939-1945. We extend this approach by considering term's immediate contexts on the co-occurrence graphs of terms. In the next step, we define temporal features of terms in order to select discriminatory terms which should be most helpful in estimating focus time of arbitrary text. Finally, the estimation of document focus time is done by extrapolating from the term focus time to the document one through a set of combination methods. Essentially, the way to measure the document focus time is to find synchronicity between different temporal pointers in the text.

We note that fundamentally our approach does not require occurrence of any temporal expressions in text. Thus documents without any explicit mentions of dates in their content can still have their focus times estimated. However, for achieving better accuracy and for the sake of completeness, we also introduce a generic method that extends the basic approach by utilizing temporal expressions such as explicit dates occurring in text.

The remainder of this paper is as follows. In the next section we review the related work. Section 3 describes methodology for the calculation of document focus time. Next, Section 4 contains the results of experimental evaluation. Section 5 provides discussion. We conclude the paper in Section 6.

2. RELATED WORK

Temporal Information Retrieval (T-IR) [1,2,4,5,9,7,14,18,19] is a subdivision of Information Retrieval (IR). It attempts to satisfy user information need by considering not only relevance but also temporal correspondence based on the underlying temporal factor behind search intent.

The usual approach to T-IR is to either use document metadata (timestamp) or to extract explicit temporal expressions from text.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505655

The problems with the first approach are as follows. First, the document timestamp is only a poor approximation of its temporal focus. Documents created, for example, in 1996 do not necessarily concern events in 1996. Second, the document timestamp is not always available. For example, web documents often lack explicit timestamps or the provided one cannot be trusted (e.g., "last-modified date" in web servers). The second approach that uses temporal expressions such as dates in text has also some problems and requires several assumptions. The current solutions use only temporal expressions occurring in text, which may be missing. Hence, for example, for query "Olympics 1964" only documents containing both "Olympics" and "1964" can be returned. Furthermore, an appearance of a date(s) does not necessarily mean that the document content is actually about the events that occurred in this date(s), as the date(s) may only be passing mentions of something weakly related to document's theme.

The closest work to ours is the one on identifying temporal intent of queries [5,14]. Metzler *et al.* [14] 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 [5] demonstrated temporal similarity measure called *GenTempEval* that associates relevant date(s) to a given query while filtering out irrelevant ones based on corpus statistics rather than document's temporal context features. There are several important differences between these works and ours. First, we work on documents instead of queries. Second, we employ more diverse range of factors (e.g., temporal entropy/kurtosis, context-based word-time associations, semantic weights and so on) and, lastly, we use news article collections as underlying knowledge bases instead of query logs or web snippets.

The research of document age estimation [8,10] can be also considered relevant to this work. De Jong *et al.* [8] and Kanhabua and Nørvåg [10] proposed temporal language models for document dating based on collections of time-stamped documents. However, as mentioned before, creation date is orthogonal to the concept of document focus time, as documents may refer to time periods different from their creation time.

Another category of works focuses on temporal information extraction from text collections. GuTime[1] and Stanford Named Entity Recognizer[2] are examples of taggers for finding dates and other temporal expressions in text. Based on temporal expression extraction, more complex systems can be built. For example, Strötgen and Gertz [19] demonstrated a system for extraction, querying, storage, and exploration of spatio-temporal information stored in text documents. Strötgen *et al.* [19] identify top relevant temporal expressions in documents either in general or with respect to a query. However, these systems rely on explicit dates and other temporal expressions in input text which, as mentioned above, may be sparse or may be missing from documents.

3. ESTIMATING FOCUS TIME

Naturally, not every document has temporal character. For example, a document explaining how to calculate integrals, or a document describing someone's house may have little to do with time. Thus we first define a notion of temporal document.

Def. 1. ***Temporal document*** is *a document whose content is related to time and which can be positioned on timeline.*

Then we define the focus time of a temporal document.

Def. 2. *A temporal document d has* ***focus time*** $\boldsymbol{\tau}$ *if the content of d refers to* $\boldsymbol{\tau}$.

Below we describe the way to estimate the focus time of temporal documents using external knowledge bases. Due to the space limitation we skip the description of the method for categorizing documents into temporal and atemporal.

3.1 Calculating Word-Time Associations

First, we calculate associations of a word with different years. They will be later used for determining document-year associations. For determining word-time associations we utilize an external knowledge base which contains references to the past associated with absolute dates. In particular, we use large datasets of news articles. Although news articles describe ongoing events, they also frequently refer to the past for variety of reasons, such as providing background, explaining the current state, comparing current events to similar ones in the past, analyzing precedence, and so on. In addition, thanks to factual and accurate reporting style, they often contain concrete dates, names as well as descriptions of past entities. Moreover, reported events are usually associated with indications of particular geographical areas in which they occurred, making it relatively easy to build collections of references to histories of particular countries.

Based on the news article collections we construct a weighted, undirected graph *G(V,E)*, where *V* denotes a set of vertices representing unique words, while *E* is a set of edges which denote word relationships. The relationships are represented as word co-occurrence relations and their weights are calculated using variation of Jaccard Coefficient [13]:

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

$c(w_i,w_j)$ is the count of sentences where words w_i and w_j co-occur together, while $c(w_i)$ and $c(w_j)$ are counts of sentences containing w_i and w_j, respectively. We use counts of sentences instead of documents as we found the sentence-based approach to perform better.

Note that we treat dates occurring in news articles as words, too. Thus by using the co-occurrence matrix we can already determine association of an arbitrary word *w* with an arbitrary time point *t* which indicates a particular year[3]. We will call such association a *direct association* and denote it as $A_{dir}(w,t)$.

Calculating word-time association in a direct way may however result in sparse results due to relatively small number of dates when compared to the number of words in documents. Therefore, we extend it by considering the word's context, that is, other words that strongly co-occur with a given word. This is reminiscent of approaches to measure semantic similarity of terms based on their contexts [13]. The intuition here is as follows.

Word w is strongly associated with time point t if many other words that strongly co-occur with w are also strongly associated with t.

This intuition is based on the assumption that a word about a given past event tends to co-occur with other words related to that

[1] http://timeml.org/site/tarsqi/modules/gutime/

[2] http://www-nlp.stanford.edu/software/CRF-NER.shtml

[3] In the experiments we assume yearly granularity, hence, we will use "time point" and "year" interchangeably throughout the paper.

event. The formula for *context-based association*, $A_{con}(w,t)$, is as follows.

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

Taking squares of the values of $A_{dir}(w_j,w_i)$ decreases the impact of terms which are weakly associated with the target word w_i. We have also experimented with the method that uses non-squared $A_{dir}(w_j,w_i)$ but we found it results in inferior performance.

Lastly, we normalize the above-described association scores of words with each time point by dividing them by the geometric mean of the association scores of all words with this time point.

3.2 Estimating Temporal Weights

Using the estimated word-time associations we next categorize words according to their discriminative capabilities for determining document focus time. We put forward the following hypothesis:

A word has high discriminative capability for determining document focus time if it has strong association with only few time points and weak association with other time points.

Thus words useful for determining document focus time should be strongly pointing to one or only few time points. In order to rank terms according to their discriminative power we propose two approaches. In the first one, we calculate *temporal entropy* of a word, which is defined as entropy over the association scores of the word with all the time points. Entropy was also used in [8] for the purpose of document dating. To calculate temporal entropy, we first normalize the association scores to obtain the probability distribution over time. Let $P_w(t_i)$ represent probability that a word w is associated with the time point t_i. Temporal entropy is then defined as:

$$E_w=-\sum_i P_w(t_i)\ln P_w(t_i) \quad (3)$$

Temporal entropy favors terms characterized by non-uniform probability distribution of their associations with time. A term with strongly varying distribution such as an "earthquake" or "war" would then have higher score than "smoother" terms like "city" or "person". However, this measure does not consider the distance between "peaks" in word-time associations. In result, temporally ambiguous terms (e.g., "earthquake" or "war") that have long distances between their peaks in the word-time probability distribution would be highly scored. Although such terms are more useful than "smooth" terms, relying on them could hinder the performance of focus time estimation due to the confusions these terms may bring (e.g., information on different earthquakes or wars at distant time points). It is thus better to find terms which have strong associations with a few nearby years or only one year. Thus, as a second approach, we propose to use *temporal kurtosis* which applies the kurtosis measure on word-time association scores. It is defined as:

$$K_w=\frac{\sum_i\left(A(w,t_i)-\mu\right)^4}{N\sigma^4} \quad (4)$$

N is the total number of time points under analysis (i.e., timeline length), while μ and σ denote the mean and standard deviation of the associations of w with the time points, respectively. $A(w,t_i)$ stands for either of word-time association measures introduced in Section 3.1.

We employ both the temporal entropy and temporal kurtosis measures to compute term weights, which will be used later in calculating document focus time. We denote temporal term weights based on the temporal entropy and kurtosis measures as ω_w^{temE} and ω_w^{temK}, respectively, while ω_w^{tem} denotes either of the weights. The values of temporal term weights are defined as follows.

$$\begin{aligned}\varpi_w^{temE}&=\max_j(E_j)-E_w\\ \varpi_w^{temK}&=K_w\end{aligned} \quad (5)$$

3.3 Calculating Focus Time

In this section we describe several approaches to compute document focus time based on the previously introduced word-time associations and temporal weights of words.

3.3.1 Calculating Document-Time Associations

First, we estimate the association of a target document with time. For this we propose the following hypothesis:

If a document d contains many words that are strongly associated with a time point t, then d has strong association with t.

Document-time association is then based on averaging time associations of terms contained in a document using their temporal weights. A simple way to do so is to use all unique words in the document.

$$S_U(d,t)=\frac{1}{|d|}\sum_{w\in d}\varpi_w^{tem}A(w,t) \quad (6)$$

$S_U(d,t)$ indicates here the association score of d with a time point t based on all unique terms in content of d, while $|d|$ denotes the number of unique terms. The score can be considered as the "goodness" of the time point t to form the focus time of d. Same as in Equation 4, $A(w,t)$ represents either of the two word-time association methods introduced in Section 3.1.

A natural extension of the straightforward approach that uses only the presence of words is a method that considers term frequency in text, similarly, to the relevance-frequency hypothesis in IR.

$$S_{TF}(d,t)=\frac{1}{\sum_{w\in d}N(w,d)}\sum_{w\in d}\varpi_w^{tem}N(w,d)A(w,t) \quad (7)$$

Here $N(w,d)$ is a function returning the number of times a word w occurs in a document d.

The intuition behind using frequency of terms in text is that the temporality of representative, important words should be good representation of the temporality of the document itself. We note however that the term frequency alone may not always be the best way for capturing the importance of terms in text. Frequent words such as stop words may carry little meaning and be thus poorly descriptive of the document content. Also, in texts about past the names of past entities or other useful clue words may sometimes appear sparsely despite being crucial for estimating the relevant time periods of documents. We thus propose extending the above approach by introducing additional weights that would represent word importance in text. In particular, we estimate the prestige value of a word in its document which is derived by recursive calculation over word interconnections in the document. The approach is similar to the one used in TextRank algorithm [16]. We calculate TextRank score of each word in a document and use its normalized versions as word's importance weight, ω_w^{imp}.

Equation 8 represents the document-time association scores based on both importance as well as temporal weights.

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

Note that the time points assigned to a document using Equations 6, 7 and 8 depend on the consensus between the associations of document terms to time. A document with many words that point to the same time point will thus have strong association to this time point. For calculating this consensus, terms with high temporal discriminative power are preferred (thanks to using temporal weights) as well as terms that are representative for the document content.

We also note that the three methods described in this section do not explicitly use temporal expressions occurring in the content of a target document, although, obviously, they constitute important evidence of a document's focus time. We then also introduce the extended approach to incorporate temporal expressions as well. For this we first extract dates from document content and then apply Gaussian Kernel Density Estimate [20] based on the extracted dates. This procedure generates mixture of Gaussian distributions with their means centred at the extracted dates. The sum of such distributions constitutes document-time association that is constructed solely on extracted temporal expressions. Let $S_{DATE}(d,t)$ denote the association score of a time point t with document d calculated in this way. Then the extended document-time association, $S_{EXT}(d,t)$, is as follows.

$$S_{EXT}(d,t) = S_{DATE}(d,t) + S(d,t) \quad (9)$$

$S(d,t)$ stands here for a document-time association score according to any of the previously described measures (Equations 6, 7 and 8).

As the last step we apply smoothing to the document-time association plots. We use here again Gaussian Kernel Density Estimate which generates and mixes Gaussian distributions positioned at every time point.

3.3.2 Computing Document Focus Time

As a simple implementation, in this proposal, we choose a single time point with the highest association score as the estimated focus time and denote it as $t_{foc}(d)$.

$$t_{foc}(d) = \arg\max_t S(d,t) \quad (10)$$

$S(d,t)$ indicates the association score between the document d and time point t according to any of the four measures described in Section 3.3.1 (Equations 6, 7, 8 and 9).

Representing the focus time of a document by a single most related time point may be preferred in applications for which the storage and processing requirements play a crucial role. For example, storing single time instances in an inverted index should have little effect on the total index size. Also calculating temporal similarity between a query and document should be relatively simple. On the other hand, it is obvious that representing focus time as a single most relevant year may not be accurate enough for certain applications. In the future we plan to propose estimating focus time as set of time periods to satisfy the requirements of more precise temporal calculation.

4. EXPERIMENTS

4.1 Constructing Knowledge Bases

First, we need to gather sufficient amount of temporally grounded historical references in order to calculate word-time association scores. For this we collected news articles published in the period of 1990-2010 from Google News Archive[4]. News articles were obtained by issuing country names as queries to the news article search engine. We used country names rather than arbitrary queries because most of the events physically take place in particular countries and such events can have diverse nature. In order to diversify datasets we decided to focus on five countries: Germany, UK[5], France, Japan and Israel.

For each country, we gathered all the returned search results with links to original articles, and downloaded their content. We discarded articles written in languages other than English using text categorization based on n-gram matching [6]. Finally, we formed 5 news article collections (each for a single country), which in total contained 535k news articles (Germany: 87k, UK: 149k, France 110k, Japan: 97k and Israel: 92k). We then extracted the core part of the news articles by removing markup and by identifying the largest chunk of text in each article. The remaining text after discarding stop words and rare terms was then used for constructing co-occurrence graphs for each country.

Next we selected temporal expressions using regular pattern expressions. We used here yearly time granularity and we also skipped relative and implicit temporal expressions as resolving them is still difficult and prone to errors [12]. The time frame of the temporal expressions was set to 1900-2013. Thus we selected only dates that fall within this time. We decided to use the above time frame as it is long enough to cover many important historical events that happened in the selected countries.

4.2 Experimental Settings

4.2.1 Datasets

For the experiments we have prepared 15 datasets grouped into 3 categories: Wiki datasets, Web datasets and Book datasets. Each category contained 5 sub-datasets, one for each different country. Table 1 shows their aggregate statistics.

Wiki datasets. We collected 250 articles from English Wikipedia[6] about major historical events related to the selected countries (50 articles for each country) which occurred within the time frame of 1900-2013. These included major wars, battles, signed treaties, strikes, elections and any other kinds of key events that we could find for the selected countries. For processing Wikipedia articles we first used the CLIPS pattern library[7] and then extracted core content of the articles removing boilerplate and references.

Wikipedia articles on past events constitute good source for the evaluation purpose as they contain precise metadata - the beginning and ending dates of each described event. Thus as a ground truth data we used information in infoboxes on the duration of the events. This data was collected manually to ensure its correctness.

Compared to other two dataset categories that we prepared, the Wiki datasets contain relatively long documents (on average, 179 sentences) with relatively many dates appearing in their content (on average, 14.5 dates in an article).

Book datasets. To construct these datasets we have used two books: "Timeline of World History" [11] and "Timelines of

[4] http://news.google.com/archivesearch

[5] Using queries: "United Kingdom", "UK" and "Great Britain".

[6] http://www.wikipedia.org

[7] http://www.clips.ua.ac.be/pages/pattern

History" [17]. These were the only books we could find that describe the historical events of all selected by us countries and which are available in an electronic form. The books cover key historical events occurring in each year of the last century in the world as short paragraphs arranged by dates. As the books do not provide separate timelines for the countries that we have chosen (only a single timeline of all major events in the world), we extracted sentences containing the name of any of the 5 selected countries or their close synonyms (e.g., "Japanese" for Japan) from within each paragraph and recorded the years of the event described in the paragraph. We then merged the sentences that had identical years of described events into separate documents.

Book datasets have moderate size of the documents (on average, 43 sentences), and relatively small number of dates appearing in content (on average, 4.5 dates in a document). Lastly, they contain documents on more recent events than the Wiki and Web datasets (the average year of events is 1982).

Web datasets. We have collected 819 texts from the following popular websites that provide historical timelines of the selected countries: "History Orb"[8], "History World"[9], "BBC Timelines"[10] and "Infoplease"[11]. We treated each paragraph as a separate document and assigned to it corresponding dates taken from the paragraph's title or manually added if the title did not contain any dates. Web datasets have documents of small size (on average, 18.3 sentences) that contain small number of dates (on average, 2.4 dates).

Table 1 Datasets statistics (aggregated over all countries).

Datasets	total #doc	avr. #sent	avr. time span of events	avr. year of events	avr. #dates
Wiki	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

4.2.2 Baselines

For comparison we use two baselines.

Random baseline. This baseline randomly estimates focus time as a random year within the set time period.

Date-based baseline. This baseline utilizes only absolute dates occurring in text that are used for generating the mixture of Gaussian distributions is the same way as described in Section 3.3.1. The date-based baseline is thus same as $S_{DATE}(d,t)$.

For the baselines we calculate focus time in the same way as in the case of our proposed methods, that is, by using Equation 10. Gaussian distributions used for smoothing have standard deviations equal to 0.6.

4.2.3 Evaluation Measure

Next, we need the way to measure the effectiveness of document focus time estimation. For this we calculate the error according to the following expression.

$$e(t_{foc}) = \begin{cases} \min\{|t_b - t_{foc}|, |t_{foc} - t_e|\} & if \quad t_{foc} \notin [t_b, t_e] \\ 0 & otherwise \end{cases} \qquad (11)$$

The time period $[t_b,t_e]$ is the true focus time of a document as given by the ground truth data. The error value as expressed by Equation 11 represents the number of years between the estimated focus time point t_{foc} and the nearest boundary of $[t_b,t_e]$. The error is then the higher, the farther is t_{foc} from the ground truth time period. It is equal to 0 if t_{foc} falls within this time period.

4.3 Experimental Results

First we have tested different combinations of proposed approaches which use only document words without relying on dates that may appear in text. Since the number of possible combinations is quite high we show only the best performing ones in Table 2[12].

Table 2 Error of the best methods (the lower error, the better).

Datasets	Method	Avr. Error
Wiki	$A_{con}(w,t)$, ω_w^{temE}, $S_{TR}(d,t)$	18.3
Book	$A_{con}(w,t)$, ω_w^{temE}, $S_U(d,t)$	16.1
Web	$A_{dir}(w,t)$, ω_w^{temK}, $S_{TF}(d,t)$	20.2

Looking at Table 2 we can see that for the Wiki datasets the best performing method has, on average, difference of about 18.3 years between the estimated focus time year and the ground truth time period. This method uses context-based association, temporal entropy and TextRank. On the other hand for the Book datasets the strongest method has average error of only 16.1 years (see Table 2); while for the Web dataset the lowest error is 20.2 years. These results are quite satisfactory considering relatively long length of the time span (over 110 years), the short average length of the described events (3.4, 4.4, 1.3 years in the Wiki, Book and Web datasets, respectively) and the variety of event types for all the concerned countries. We emphasize here that the results in Table 2 do not include the approach based on dates in text, which will be discussed in the later part of this section. Hence they are based solely on content words without using any temporal expressions.

We can notice that the context-based association between words and time points, $A_{con}(w,t)$, works well and is the component of the best performing methods for the Wiki and Book datasets. Hence, using the year associations of words that are strongly co-occurring with the target word seems to help better estimate the focus time.

We can also observe that the temporal weights (temporal entropy, ω_w^{temE}, and temporal kurtosis, ω_w^{temK}) are useful for measuring document focus time. Next, we can observe that the TextRank measure is useful when document length is large such as for the case of the Wiki datasets. For the Book and Web datasets the best methods use the term frequency for document-time association, $S_{TF}(d,t)$, or the one based on unique terms, $S_U(d,t)$.

In Table 3 we compare the results of the selected combination of the proposed methods against the results generated by the baselines. For the comparison we have chosen the method ($A_{con}(w,t)$, ω_w^{temK}, $S_{TR}(d,t)$) that performed consistently well on different datasets. It uses the context-based association, temporal kurtosis as temporal weights and TextRank scores as importance weights. We also use its extended version that incorporates dates occurring in text (see Equation 9). The proposed methods will be called *Prop* and *PropExt*, respectively.

[8] http://www.historyorb.com

[9] http://www.historyworld.net

[10] http://www.bbc.co.uk/history

[11] http://www.infoplease.com

[12] The results for the methods in Table 2 are statistically different from the random baseline as measured by t-test ($p < 0.05$).

When looking at Table 3 we can observe that *Prop* achieves better results than the baselines for all the datasets except for the case when the date-based baseline is applied on the Wiki datasets. Wikipedia articles contain relatively many dates (see Table 1); hence, the straightforward date-based baseline performs better. However, as mentioned before, many texts about past may not contain any temporal expressions or may contain only few of them. Hence, as we can see, for the Book and Web datasets, the date-based baseline performs very poorly. Thus an extended approach should be preferred over the one that relies on dates only.

Indeed, we notice that the *PropExt* performs best. For example, we can see that for the Wiki datasets the error is on average only 2.83 years. However, we note that after applying the Tukey HSD test ($p < 0.05$) to the results we found the lack of significant difference between the date-based baseline and *PropExt* on the Wiki datasets. All other comparisons between the results of the baselines and the ones of the proposed methods were significantly different. This implies that when sufficiently many dates are present in text, as is in the case of the Wikipedia documents, the improvement over the date-based approach may not be significant. Anyway, the benefit due to applying the extended method *PropExt* is still significant for other datasets that have fewer dates in content.

Table 3 Comparison with baselines.

Datasets	Random	Date-based	Prop	PropExt
Wiki	36.5	3.02	18.3	**2.83**
Book	39.3	48.1	23.5	**20.4**
Web	40.5	53.4	23.6	**20.7**

5. DISCUSSION

Data sources and "history spaces". In our approach we assume existence of different histories such as histories of countries, regions, concepts, scientific fields, etc. Accordingly, the focus time of documents could be estimated using such diverse historical spaces. In this paper, we use country-biased news datasets. This allows positioning documents on the timelines representing the histories of the corresponding countries.

Time granularity and events without time references. Using finer granularity expressions could help to more precisely quantify document focus time making it possible to find documents on a particular month or even a day. Another issue concerns estimating focus time of documents about events which are rarely associated with any explicit dates. For example, for certain events precise starting and ending dates may not be known. Solving this problem might require special type of temporal inference.

Applications. Document focus time could be considered as a complement to semantic representation of documents. One could then calculate *temporal similarity* between two documents in parallel to their content similarity. Similarly, temporal representation of a query could be matched to the one of document. For example, one may search for documents about "Hitler" (semantic part) that cover events that occurred in 1933-1934 (focus time). Finally, when using also document creation time, one could formulate complex temporal queries (e.g., query searching for documents about Hitler that focus on 1933-1934 and were created in 1956-1960). Such generic temporal queries could be useful in temporal collections like document archives.

Besides T-IR, document focus time could be useful in other potential applications. We list some of them below:

- Improving temporal expression annotation and extraction
- Detecting diverse types of references to past in texts
- Improving ordering of sentences in automatically created document summaries
- Image dating by using focus time of surrounding texts
- Supporting computational history and culturonomics [3,15]

6. CONCLUSIONS

Time is an important aspect of text. Properly estimating content time of temporal documents would help to improve Temporal IR as well as strengthen our means of analyzing and understanding documents and temporal references in text. In this paper, we describe the concept of document focus time. We also provide a range of methods for its estimation. Our approach harnesses corpus statistics and, in particular, absolute references to past years in news articles. The important characteristic of our proposal is that it also works for documents that do not contain any temporal expressions or contain only few of them.

7. 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".

8. REFERENCES

[1] Alonso, O. et al. Temporal Information Retrieval: Challenges and Opportunities. In *TWAW 2011*, pp. 1-8

[2] Arikan, I. Bedathur, S.J. and Berberich, K. Time Will Tell: Leveraging Temporal Expressions in IR. In *WSDM 2009*

[3] Au Yeung, C.-M. and Jatowt, A., Studying how the Past is Remembered: Towards Computational History through Large Scale Text Mining. In *CIKM 2011*, pp. 1231-1240

[4] Berberich, K., Bedathur, S.J., Alonso, O. and Weikum, G. A Language Modeling Approach for Temporal Information Needs. In *ECIR 2010*, pp. 13–25, 2010

[5] Campos, R., Dias, G., Jorge, A. M., and Nunes, C. GTE: A Distributional Second-Order Co-Occurrence Approach to Improve the Identification of Top Relevant Dates. In *CIKM 2012*, 2035–2039

[6] Cavnar, W. B. and Trenkle, J. M. N-Gram-Based Text Categorization. In *SDAIR 1994*, pp. 161-175

[7] Jones, R., and Diaz, F. Temporal Profiles of Queries. In *TOIS: ACM Transactions on Information Systems*, 25(3), 2007

[8] Jong de, F.M.G. and Rode, H. and Hiemstra, D. Temporal Language Models for the Disclosure of Historical Text. In *AHC'05*, pp. 161-168

[9] Kanhabua, N., and Nørvåg, K. Determining Time of Queries for Re-ranking Search Results. In *ECDL 2010*, pp. 261-272, 2010

[10] Kanhabua, N., and Nørvåg, K. Using Temporal Language Models for Document Dating, In *MLKDD 2009*, pp. 738-741, 2009

[11] Kerr, G. *Timeline of World History*, Canary Press, 2011

[12] Mani, I., and Wilson, G. Robust Temporal Processing of News. In *ACL 2000*, pp. 69-76, 2000

[13] Manning, C. and Schütze, H. *Foundations of Statistical Natural Language Processing*, MIT Press, 1999

[14] Metzler, D., Jones, R., Peng, F., and Zhang, R. Improving Search Relevance for Implicitly Temporal Queries. In *SIGIR 2009*, 700-701

[15] Michel, J.-B. et al. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014), pp. 176-182, 2011

[16] Mihalcea, R., and Tarau, P. Textrank: Bringing Order into Text. In *EMNLP 2004*, pp. 404-411. 2004

[17] Ratnikas, A. *Timelines of History*, 2012 (Kindle edition)

[18] Strötgen, J. and Gertz, M. TimeTrails: a system for exploring spatio-temporal information in documents. In *VLDB 2010*, pp. 1569-1572

[19] Strötgen, J. Alonso, O. and Gertz, M. Identification of top relevant temporal expressions in documents. In *TempWeb 2012*, pp. 33-40

[20] Sheather, S.J. Density Estimation. *Statistical Science*. Vol. 19, Number 4, pp. 588-597, 2004