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/© 2015 Elsevier Ltd. All rights reserved.

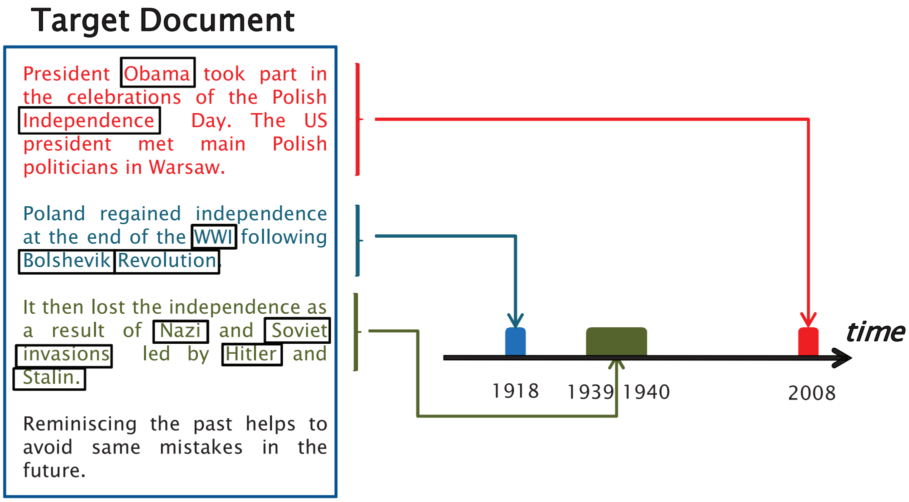
Target Document
President Obama took part in the celebrations of the Polish Independence Day. The US president met main Polish politicians in Warsaw.
Poland regained independence at the end of the WWI following Bolshevik Revolution
It then lost the independence as a result of Nazi and Soviet invasions led by Hitler and Stalin.
Reminiscing the past helps to avoid same mistakes in the future.
time
1918
1939
1940
2008

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