Document in Context of its Time (DICT): Providing Temporal Context to Support Analysis of Past Documents

Adam Jatowt[1], Ricardo Campos[2], Sourav S Bhowmick[3] and Antoine Doucet[4]

[1]Kyoto University & AIRC, AIST
jatowt@gmail.com

[2]Ci2-Smart Cities Research Center - Polytechnic Institute of Tomar, LIAAD–INESC TEC
ricardo.campos@ipt.pt

[3]Nanyang Technological University
assourav@ntu.edu.sg

[4]University of La Rochelle
antoine.doucet@univ-lr.fr

ABSTRACT

Old documents tend to be difficult to be analyzed and understood, not only for average users but oftentimes for professionals as well. This is due to the context shift, vocabulary evolution and, in general, the lack of precise knowledge about the writing styles in the past. We propose a concept of positioning *document in the context of its time*, and develop an interactive system to support such an objective. Our system helps users to know whether the vocabulary used by an author in the past were frequent at the time of text creation, whether the author used anachronisms or neologisms, and so on. It also enables detecting terms in text that underwent considerable semantic change and provides more information on the nature of such change. Overall, the proposed tool offers additional knowledge on the writing style and vocabulary choice in documents by drawing from data collected at the time of their creation or at other user-specified time.

Keywords

Document analysis, historical texts, document archives, language evolution, semantic word change

ACM Reference format:

Adam Jatowt, Ricardo Campos, Sourav S Bhowmick, & Antoine Doucet. 2019. Document in Context of its Time (DICT): Providing Temporal Context to Support Analysis of Past Documents. In *Proceedings of ACM CIKM Conference (CIKM'19), November 3–7, 2019, Beijing, China.* ACM. NY, NY, USA 4 pages. https://doi.org/10.1145/3357384.3357844

1. INTRODUCTION

In recent years, frequent initiatives aimed at digitalization of historical texts were carried out by memory institutions like libraries, museums, and state or national archives. Old books, news articles, letters, legal documents, and other document types have been then made publicly available as large open collections (e.g., Project Gutenberg[1] or Internet Archive Text Collection[2]). Certain professionals and experts need to work with such documents analyzing them for variety of reasons. For example, humanists investigating old literature or historians trying to find connections between historical events spend considerable time studying in detail the writings of the past.

When investigating such archival documents, a present-day user implicitly takes the viewpoint of the current time. Yet, to correctly understand documents or document sub-collections (e.g., legacy of the same historical author), we need to set them in the temporal context of their times. That is, the usage of words, the style of writing and other aspects of the text need to be seen in relation to ones of the contemporary times when the text was written. While this may be possible for some users, average users and even professionals alike may lack such skills.

In this work, we demonstrate a novel system for investigating content of documents (typically, born-analog documents that were subject to digitalization and optical character recognition (OCR)) by reference to the state of language used in the time when the documents were created. It is not immediately obvious whether the style of writing and vocabulary choice made by the author were frequent and whether they were used previously for a long time or rather were novel at the time of the text creation. Such information, if provided, would then shed new light on the writing style and the relation of a given vocabulary to its use frequency at the time of the document creation. In particular, in the proposed system we indicate the frequency of words or n-grams in text in relation to how language was used in the past.

Additionally, the system highlights terms in text that underwent considerable semantic change. It is not a secret that old documents tend to be difficult to read and understand, not only for average users but even for experts, too. To large degree this is because many words shifted their meaning over time [5,7,9]. The word coaches, for instance, was used at the beginning of the 20th century to refer to cars, while nowadays it is mostly used in the context of sports to refer to trainers. Detecting this kind of terms and providing more information on the nature of their change would help users and professionals to better comprehend the overall meaning and characteristics of texts. All these cases are targeted in the proposed demonstration system which provides contextual temporal knowledge, in a visual and interactive way about input documents based on the associated long-term corpus.

To sum up, we propose a novel way of investigating old documents by seeing them through the lens of their (or even other) times. We call it a *Document in Context of its Time* (DICT), in a loose parallel to the well-known Key Word in Context (KWIC) viewing and analysis approach. We also design a proof-of-concept system for supporting such analysis and demonstrate its functionalities: *n-gram frequency analysis*, *n-gram age analysis*, *semantic change analysis* and *term-oriented interactions*. Our system could help to improve understanding of past documents and could also support designing better methods for timestamping documents [2,4,6] or help with post-OCR error recognition.

[1] http://www.gutenberg.org
[2] http://www.archive.org

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3357844

2. RELATED WORK

Our proposal resonates well with the notion of *historical contextualization* - a fundamental concept behind historical thinking and practice, which involves the ability to put something in its proper historical context and understanding an event or a document in relation to what else was happening at the same time. The only related work for historical contextualization from the computational aspect, that we are aware of, is about Learning-to-Rank based approach for finding Wikipedia abstracts that would help to clarify the nuanced meaning of text passages in target past documents [10] and a related demonstration web system [11].

Language change has been of interest in linguistics for quite some time [7]. In particular, tracking evolving semantics of words has garnered quite much attention of both scientists as well as of the wider public. Recently, computational approaches have been frequently employed for supporting the analysis of diachronic word change [9,3]. Most of the approaches take a query word and output its change points and/or set of different senses the word assumed over time. Our proposal integrates document analysis and semantic change analysis, as well as data on vocabulary use for comprehending old documents.

Research on automatic document dating is also related to our work [2,4,6]. Typically, large scale diachronic datasets are used for this task to construct historical language models that would support document creation date inference. In our previous work we have demonstrated an interactive, related system that permits document age estimation by aggregating time series plots of its n-grams [2]. The current work has different focus of supporting archival document comprehension and analysis from diverse angles based on large scale historical accounts of language use.

3. DATASET & PREPROCESSING

To accomplish our objectives, we need a dataset which is large enough for drawing valid conclusions for quite a long span of time. To this end, we utilize *Google Books N-grams*[3] - a compilation stemming from the Google Books project which claims to have processed data from about 6% of ever published books. Google Books N-gram datasets have been used for *culturonomics* [8] - a study of the changes in word usage and cultural trends over time as well as have been increasingly employed for computational approaches towards diachronic word analysis [9]. In this work, we use *Google Books 5-grams* for reasoning about temporal characteristics of documents. We believe that due to its large size, these datasets are the most appropriate resource for representing word use across time. We also note that the data are provided not only for English but also for Chinese (simplified), French, German, Hebrew, Italian, Russian, Spanish as well as in some cases it has been derived from specialized English corpora, such as American English, British English, English Fiction.

To smoothen the time series plots and to provide more intuitive and easy operations, the data have been integrated into decades and most of the time used as such. We also normalized n-gram plots by the total data size at each decade. Furthermore, to remove tokens generated as a result of OCR errors or those specific only to a particular document or an author, we applied a threshold for removing rare words which was set to 300 words per each decade. Finally, a database containing the processed data for different n-gram sizes has been created where each individual n-gram is associated with its frequency plot over the entire time span.

[3] http://books.google.com/n-grams/datasets

4. SYSTEM DESCRIPTION

Our system requires an input text to be pasted in the main text form. It then generates a series of heat maps laid over the input content based on the time series plots of extracted n-grams and the semantic representation of the contained terms. Most of these views are determined based on a document timestamp (which is assumed to be known precisely or at least approximately).

Besides inputting text, the user sets also the time range using the time slider to limit the scope of analysis. This is useful when one wishes to analyze more closely the data over a particular sub-period. Also, the data tends to be relatively sparse for early decades, such as ones before the 18th century, so sometimes it might be good to constrain the temporal range. The next parameters to set up are n which is the number of grams to be considered ($1 \leq n \leq 5$, where n=1 means the level of individual token) and θ - a threshold parameter on n-gram frequency for estimating start and end dates of word use and for determining neologisms/anachronisms. Other possible options let users choose if the word case and punctuations are to be considered during n-gram extraction and matching.

During the execution, n-grams of a given size n are extracted from the input text and matched to the underlying database. This is done by employing a sliding window(s) of length n over the input text. Each n-gram found in the text is then searched in the database for collecting its time series data.

In the following we will describe 4 integral components of the proposed system: *frequency analysis*, *age analysis*, *semantic change analysis* and additional interaction functionalities that enrich these components.

4.1 Frequency Analysis

In the first view (exemplified in Fig. 1) we show the degree to which single words or n-grams in a document were used in the past. The redder the background color in this view, the more often the word was used at a certain time (e.g., document creation or publication time which is set by a user). For the case of $n > 1$, the color of a word is decided based on the aggregated frequency of n-grams covering this word.

Using the frequency view one can immediately spot which n-grams were rare or unique during the document creation date (this date needs to be manually entered by the user using a time slider at the top of Fig. 1). For example, the 3-gram "was not that" was quite common in 1890s, while "akin to love" was rare as it can be seen in Fig. 1.

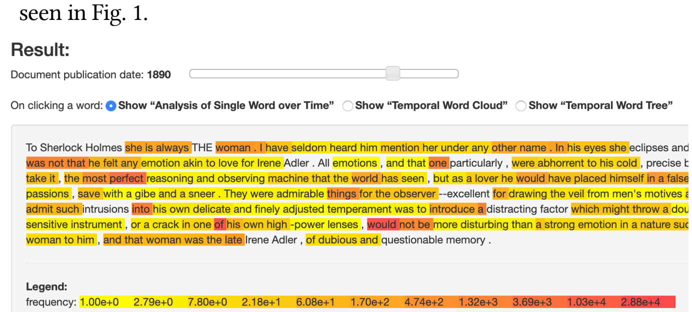

Figure 1. Heatmap view indicating term frequency using 3-grams of an example document (excerpt from "A Scandal in Bohemia" book by Arthur Conan Doyle) at its creation date: 1890. Image has been truncated for saving space.

As an interactive option, the system allows setting the time in a dynamic fashion such that it is possible to investigate the changes in the frequency view diachronically when sliding the time bar towards or far away from the present. When setting the current decade one can, for instance, notice n-grams that are rare for present-day readers. As another example, Fig. 2 shows the same text, yet with a different assumed publication date (1974). Focusing on rare words, one could investigate if they would have remained rare, had the document been created at another date.

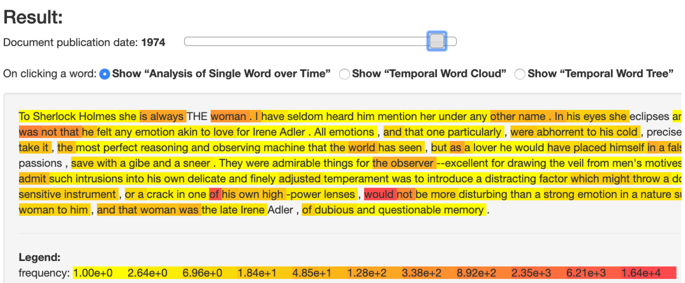

Figure 2. Heatmap view indicating term frequency in the same document and settings as the one in Fig. 1 at another date: 1974. Image has been truncated for saving space.

4.2 Age Analysis

In the next mode, called *age analysis*, a user can analyze the age of n-grams used in document content. In particular, he or she can find whether terms in the document were new or rather old at the time of its creation, that is, if they have been used for a relatively short or rather long time until document creation date based on *Google Books 5-grams* dataset. For example, when studying a book, it is possible to learn in which contexts the book's author used neologisms and in which rather old and well-established terms.

We compute it as follows: For each extracted n-gram in text the system finds in its frequency plot the *oldest decade* since when the n-gram has been used in the past. To conveniently judge the first use of the n-gram, we set up a threshold parameter θ on the normalized frequency plot of the n-gram. Hence, the oldest decade is the first one when the n-gram frequency reached a value higher than θ. Based on this, in the *n-gram age view*, the color of each n-gram represents the oldest decade of the n-gram (see Fig. 3 for example of 1-gram analysis). The older this decade, the redder the background color of the n-gram should be. On the other hand, the younger the decade, the yellower the background color is.

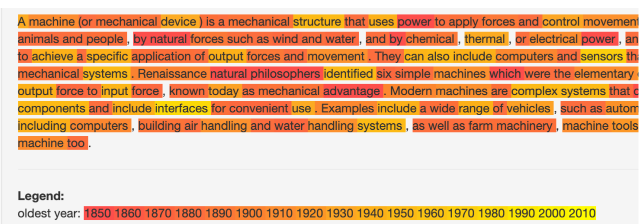

Figure 3. Heatmap of *1-gram age view* of an excerpt from Wikipedia page on machine [4]. Red colors indicate the increasing age of terms. Image has been truncated for saving space.

For example, based on the view in Fig. 3, "power" is an old term (existing since before 1850 as the time period was capped) while sensors is not so much. We note that additionally, the system allows to indicate anachronisms in text, which are n-grams that were not used anymore in language at a particular date. In practice it means estimating the *latest decade* of an n-gram in a parallel fashion to the estimation of the oldest decade (finding the decade after which the n-gram was never used anymore with the frequency over θ).

Additionally, based on the *oldest decade* information, the system allows answering question regarding the knowledge of n-grams by past readers if a document would be published at any hypothetical date. For example, the same content from Fig. 3 is now shown in Fig. 4 as if it would had been created, for example, in 1869. Based on grey background coloring we can see which words would not have been known by the readers reading the document at that year. Such an option could support document age estimation and could help testing assumptions on document creation date based on the interplay between terms' usage in language and the assumed creation date.

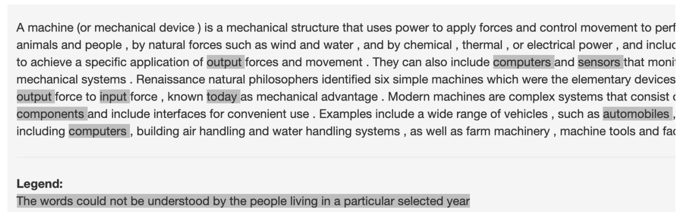

Figure 4. Indication of unknown terms of an excerpt from Wikipedia page about "machine" should it be created in 1869. Grey color indicates the terms not used in 1869. Image has been truncated for saving space.

4.3 Semantic Change Analysis

All the above-discussed views are based on frequency information. In the next view we refer to word semantics. In particular, we capture data related to diachronic evolution of words contained in text for constructing a unified view of semantic change.

For representing the word meaning we adopt a common approach used in NLP, the distributional semantics [1], based on which a word's meaning is captured by co-occurring words (hereafter called context terms). For a given target word w[5] in a decade d, we collect all n-grams that contain w. We then sum the counts of all context terms. The word representation in d is then given by a vector, whose size is the number of unique words found in the dataset, while the weights are calculated as the normalized counts of context terms co-occurring with w in d.

In this view, the system evaluates the degree of each word's context change across time. In particular, the vector representation of the target word at a reference decade d_r (typically the decade denoting the document creation time) is compared with the one of the current decade d_{now} for capturing the degree of word's semantic change. If the similarity between the word's vector at decade d_{now} and the one at reference decade d_r is low (i.e. $sim(d_{now}[w], d_r[w]) \rightarrow 0$), then a semantic change is likely to have occurred between these two decades. We use cosine similarity as the measure of context similarity, with an option to remove stopwords. Note that the choice of a method behind the word semantic change estimation is orthogonal to our system and other more refined solutions like [5] (or others discussed in [9]) can be applied instead. In the current implementation we use a simple solution that can be also easily explained to professionals outside of computer science.

Fig. 5 shows the heatmap view in this mode where the semantic change degree is represented on a color range from red (large change) to blue (no or small change). Words that underwent the

[4] https://en.wikipedia.org/wiki/Machine (note that we chose a present document here for the convenience of explanation).

[5] For simplicity, we provide semantic change analysis for words only.

largest level of the semantic change (e.g., finely or cold in Fig. 5) may be least understandable to current readers. Therefore, the proposed view may also serve as an indicator of readability issues/difficulties that present-day readers may encounter when viewing the documents. Understandingly, we noticed that named entities are often indicated as having undergone large change (e.g., different persons but same names across time).

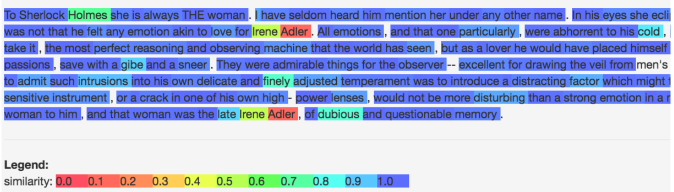

Figure 5. Semantic change view of an excerpt from "A Scandal in Bohemia" book by Arthur Conan Doyle published in 1890. Image has been truncated for space.

4.4 Word-centered Interactions

The system allows also for several investigations based on the above-discussed views. First, in the frequency view (Sec. 4.1) and in age analysis view (Sec 4.2), clicking on each word shows the pop-up window with the frequency plots over time of its covering n-grams (see Fig. 6 for example). In the age analysis view, 's value is also shown against the plot, and the estimated oldest as well as the latest decades of covering n-grams are listed.

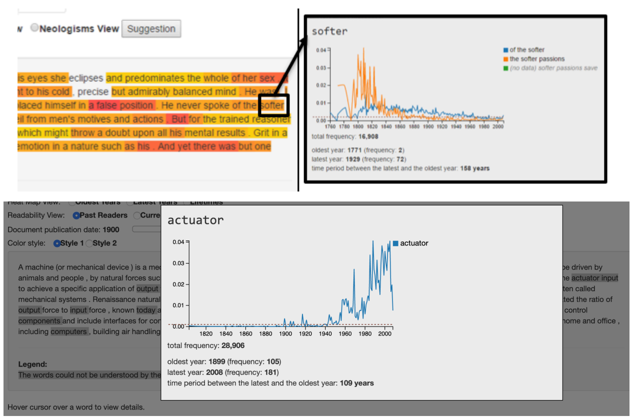

Figure 6. Examples of investigation of word frequency in the frequency view (top) and age analysis view (bottom).

For the semantic change analysis (Sec. 4.3), clicking on a word outputs self-similarity plots of the word based on comparisons of its contextual representations in different decades. Fig. 7 demonstrates the result obtained for the word cry from a book that was published in 1915. The top graph shows by the thick blue line the similarity plot between the word's semantics in the reference decade (2000s) and the ones in each past decade using cosine similarity. It permits understanding when the meaning of the word became similar over time to its meaning in the reference decade. The curve with a high and steep increase would characterize a word which acquired the present meaning in a relatively short time period. On the other hand, a flat curve indicates words with stable or slowly changing meaning over time. The bottom plot in Fig. 7 adds more evidence to the word evolution analysis by outputting similarities between each pair of consecutive decades. It is then possible to examine how the word changed from decade to decade.

The last interaction way supports investigating change in the word's context over time. The previously described views do not permit investigating term sequences. Showing the order of context terms as for how they appeared together with the target selected word can however provide novel insights to support broad investigation of word change over time (e.g., typical preceding and following words at different decades). To reflect word sequences, we employ word tree technique [12] - a visualization style as well as an information-retrieval technique for rapid exploration of large bodies of text. However, the word tree is designed for single texts or synchronic document collections and has not been explicitly used to long-term diachronic document collections as in our case. We then adapt it to provide time-based word order visualization (the image is not shown due to space limitation but can be provided upon request). The term size reflects term frequency at a given relative position after or before the queried word, same as in usual visualization. However, to convey temporal information, the system also displays the frequency plot of each term sequence together with the plots of its extended sequences under each term. Sequence extension is done by gradually appending the terms following the previous sequence. Thus, along with the order-related information, the user also receives temporal information related to each particular term sequence. Note that the frequency plot of a given sequence subsumes all the frequency plots of its extended sequences. Since the number of total words grows very fast with each consecutive position we provide an option to limit the number of words displayed at each position (e.g., top 5 words).

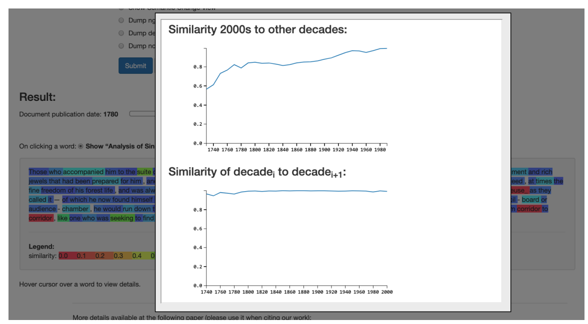

Figure 7. Semantic change of clicked word "cry" in the phrase "spoke of the cry of pleasure" from the 7th edition of "A house of Pomegranates" by O. Wilde published in 1915.

Acknowledgments. This work is financed in part by JSPS Kakenhi grants (#17H01828, #18K19841) and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project: UID/EEA/50014/2019.

5. REFERENCES

[1] Z. Harris, Distributional Structure. *Word*, 10(23):146–162, 1954.
[2] A. Jatowt and R. Campos: Interactive System for Reasoning about Document Age, In CIKM 2017, 2471-2474, 2017.
[3] A. Jatowt, R. Campos, S. S Bhowmick, N. Tahmasebi and A. Doucet: Every Word has its History: Interactive Exploration and Visualization of Word Sense Evolution, In *CIKM 2018*, 1899-1902
[4] N. Kanhabua, K. Nørvag. Improving Temporal Language Models for Determining Time of non-timestamped Documents. In *ECDL 2008*, 358-370.
[5] V. Kulkarni, R. Al-Rfou, B. Perozzi, S. Skiena. Statistically Significant Detection of Linguistic Change. In *WWW 2015*, 625-635.
[6] A. Kumar, M. Lease, J. Baldridge. Supervised Language Modeling for Temporal Resolution of Texts. In *CIKM 2011*, 2069-2072.
[7] W. Labov. *Principles of Linguistic Change (Social Factors)*, Wiley-Blackwell, 2010.
[8] J.-B. Michel et al. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014), 176-182, 2011.
[9] N. Tahmasebi, L. Borin, A. Jatowt: Survey of Computational Approaches to Lexical Semantic Change. *CoRR abs/1811.06278*, '18
[10] N. K. Tran, A. Ceroni, N. Kanhabua, and C. Niederée: Back to the Past: Supporting Interpretations of Forgotten Stories by Time-aware Re-Contextualization. In *WSDM 2015*, pp. 339-348, 2015.
[11] N. K. Tran, A. Ceroni, N. Kanhabua, and C. Niederée: Time-travel translator: Automatically contextualizing news articles. In *WWW2015*, pp. 247-250. 2015.
[12] M. Wattenberg, F.B. Viégas, The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on* Visualization and Computer Graphics (2008): 1221-1228.