Time-centric Exploration of Court Documents

Philip Hausner
Institute of Computer Science
Heidelberg University, Germany
hausner@stud.uni-heidelberg.de

Dennis Aumiller
Institute of Computer Science
Heidelberg University, Germany
aumiller@informatik.uni-heidelberg.de

Michael Gertz
Institute of Computer Science
Heidelberg University, Germany
gertz@informatik.uni-heidelberg.de

Abstract

Getting an overview of a complex phenomenon that is described in numerous documents poses a major challenge in many application domains, among which the legal domain is of particular societal interest. In this paper, we outline a framework that is based on constructing term co-occurrence networks from documents and that allows users to explore a collection of court documents in a time-centric fashion, thus providing insights into a case's chronology and entities involved.

1 Introduction

Lawyers and judges are often facing complex court cases that comprise hundreds of documents that cover charges, expert opinions, witness accounts, and the like. Prominent examples are well known in the context of the Enron scandal [Wik20b], the Panama papers [Wik20d], the Cum-Ex-Files [Wik20a], or the National Socialist Underground (NSU) trial [Wik20c]. Even though many of the documents are available in electronic form (mostly as PDFs), getting an overview of the case in terms of applicable statutory violations, relevant statutes, people and organizations involved as well as the temporal development of the case under consideration play a crucial element in the daily investigative business of a jurist.

While typical Natural Language Processing tasks such as Named Entity Recognition already provide valuable information when extracted from court documents, the organization of these concepts to present a jurist an overview and starting point for further analyses and focused reading is still a challenge. This is particularly problematic as there is no default by which documents and texts can be arranged to provide for a comprehensive reading, a problem forensic search or e-Discovery systems used in the legal domain are also facing.

In this paper, we outline a time-centric approach that aims to arrange key information from court documents using timelines in a flexible manner. The key idea is to construct weighted term/entity co-occurrence networks around temporal expressions detected in the texts. For the weighting, we introduce a TF-inverse timestamp frequency metric to score the relevance of temporal expressions, exploiting the natural time hierarchy (days,

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia (eds.): Proceedings of the Text2Story'20 Workshop, Lisbon, Portugal, 14-April-2020, published at http://ceur-ws.org

months, years). The constructed networks can be arranged along a timeline and allow for different exploration tasks, including the investigation of named entities at different points in time as well as temporally-centered zoom operations.

In the following section, we briefly outline related work. In Section 3, we detail the time-centric network model, followed by experimental results based on documents from the above-mentioned NSU trial in Section 4.

2 Related Work

Temporal information is inherent in many documents, and due to its wide variety of applications an important research subject. In information retrieval, for example, it is crucial for temporal clustering of documents or temporal question answering [ASBYG11]. Important for all these approaches is the accurate extraction and normalization of temporal expressions from textual data using state-of-the-art temporal taggers like HeidelTime, which is domain-sensitive and applicable to a wide variety of languages [SG13, SG15]. Furthermore, temporal information can be utilized for timeline summarization to give a compact overview of a topic. For example, Steen and Markert introduced an abstractive timeline summarization model that computes timelines completely unsupervised using multi-sentence-compression [SM19]. However, timelines are not widely used for exploratory tasks as can be for example seen in the survey of Campos et al. [CDJJ14]. Alonso et al. employed a timeline visualization for the exploration of search results [ABYG07], and Tuan et al. constructed timelines from Wikipedia articles and employed extracted contexts to summarize the events associated with an entity. Furthermore, Prytkova et al. introduced a similar graph model to the one employed in this paper, although they did not formalize their approach in the form of a timeline [PSW12]. In the legal domain, Knight et al. were one of the first to consider temporal information [KMN98], and Lagos et al. discuss the value of timelines for legal case building [LSCO10]. Nonetheless, to the best of our knowledge not much research about time-based data has been done in the legal domain yet. Probably most similar to our work is the model of Spitz et al. who provide a weighted bipartite graph model that is partitioned into dates and other (non-date) terms, and that can be used for temporal analysis [SSBG15]. However, in their model only the relation between dates and other terms can be observed, while oftentimes the relation between terms around a timestamp is of relevance. The model proposed in this paper aims to achieve this by introducing a separate graph for each point in time.

3 Time-centric Graph Representation

In this section, we establish a model that allows for the description of dates with the help of graphs by representing each date by its own network, employing node weights to express the importance of a term for a date. Ultimately, these graphs are utilized to construct the timeline visualizing the contents of a given document collection.

3.1 Time-Centric Graph Model

Let $\mathcal{P}$ be a collection of documents (or *pages*). Moreover, each document $p \in \mathcal{P}$ consists of a set of sentences $s \in p$, and we denote the set of all sentences with $\mathcal{S} = \cup_{p \in \mathcal{P}} \{s \mid s \in p\}$. A sentence in this model is treated as a bag of words, and while two sentences may contain identical words, they are treated as separate in this model. Additionally, some words carry temporal information, which can be extracted as dates d. The set of all dates present in the data set is denoted as D, and two dates are considered equal if they describe the same date (e.g., a year or day). Furthermore, to account for the differences in the granularity of dates, we partition D into $D = D_y \cup D_m \cup D_d$ where the indices denote years, months, and days, respectively. For this partitioning a hierarchy can be formulated, i.e., for each day exists a month in which it is included, and the same relation holds between months and years.

3.1.1 Time-centric Co-occurrence Graph

Given a set of dates D, a **time-centric co-occurrence graph** is a weighted graph $G_d = (N_d, L_d)$ with nodes N_d being the terms extracted in a window of x sentences around timestamp $d \in D$, and links L_d that represent the co-occurrences between terms in the same context around timestamp d. Since all terms in the context window around one instance of a timestamp d co-occur in this model, each subgraph extracted around a specific occurrence of a timestamp has to be fully connected. We denote the set of all time-centric co-occurrence graphs of D with $\mathcal{G}_D = \{G_d \mid d \in D\}$. For each date $d \in D$, there exists only one graph representing the date; this means in particular that there is not a separate graph for each occurrence of d, but co-occurrences around different

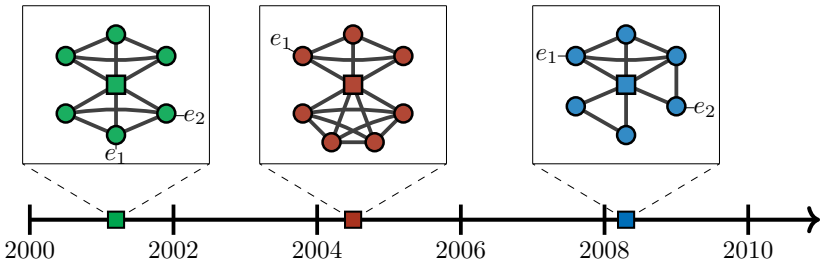


Figure 1: Timeline employing time-centric co-occurrence graphs for three different points in time (green, red, blue). Each timestamp has a graph assigned that is visualized in its corresponding box indicated by matching colour. In each graph the central node represents the respective date, the rest are co-occurring terms.

instances of d are aggregated in the same graph G_d. Taking into account the partitioning hierarchy described above, it can be stated that for each graph G_{d_1} that represents the network of a given day d_1 and contains an edge e, e also exists in graph G_{m_1} of the month m_1 containing d_1; the same holds for months and years. We additionally define a function **sent** : $L_d \rightarrow \mathcal{P}(\mathcal{S} \times \mathcal{S})$ that assigns to each edge all pairs of sentences from which it was created, i.e., in which two (not necessarily distinct) sentences the two nodes co-occurred. **sent** enables exploration of the document collection by offering a way to the user to show the relation between two terms, as well as their origin, which means their mutual co-occurrences, in the document collection.

3.1.2 Node and Edge Weightings

Nodes as well as edges of a time-centric graph are assigned a weight. Edge weights are scaled by the number of times both terms co-occurred divided by the maximum number of co-occurrences in the graph. Node weights are computed by an adaption of the tf-idf weighting scheme we call *term frequency - inverse timestamp frequency* (*tf-itf*) defined as:

$$\text{tf-itf}(n, d, D) \coloneqq \text{tf}(n, d) \cdot \text{itf}(n, D), \tag{1}$$

where tf is the number of times term n occurs in the context window around timestamp $d \in D$ normalized by the total number of words occurring in the context windows, and itf is defined as

$$\text{itf}(n, D) = log\left(\frac{M}{1 + |\{d \in D : \text{tf}(n, d) > 0\}|}\right), \tag{2}$$

with M being the number of unique timestamps in the document collection. A term has *tf-itf rank* m with regard to a time-centric co-occurrence graph, if it is the term with the m-th highest tf-itf.

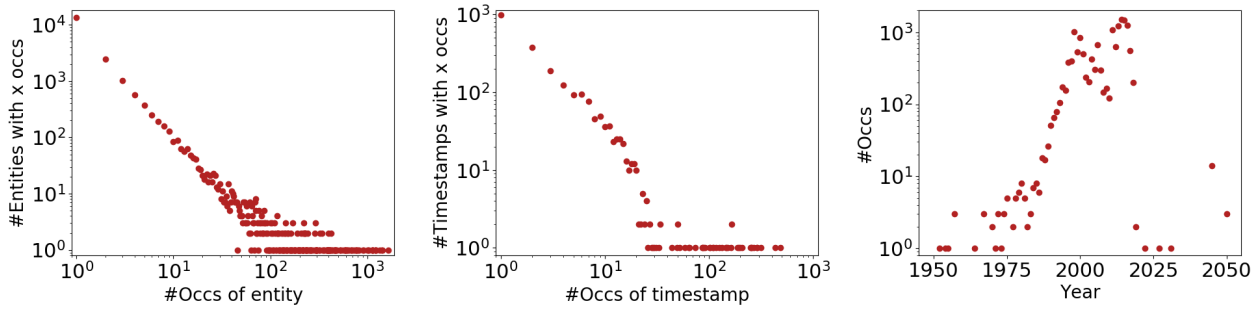
3.2 Timeline Representations

The resulting time-centric co-occurrence networks can be arranged on an appropriate timeline as indicated in Figure 1. Such a timeline can be effectively utilized for a variety of exploration scenarios. In the following, we discuss two prominent use cases: Entity-centric timelines and zooming operations. An entity in this context is a named entity, i.e., a person, a location, and the like.

3.2.1 Entity-centric Timelines

For entity-centric timelines, we do not take all timestamps into account for which a time-centric graph G_d is constructed, but only those time-centric graphs that exhibit a desirable property associated with the presence of one or more entities. While such a property can be highly complex, for data exploration tasks presented here, it is sufficient to stick with one of the two following criteria:

1. One or multiple entities E need to occur in the associated graph G_d, i.e., they are represented as a node in the network.
2. One or more edges e between certain entities have to exist in G_d, i.e., they have to be directly connected for the timestamp being part of the timeline.



(a) (b) (c)

Figure 2: Properties of the NSU court trial. (a) Log-log plot of the occurrence distribution over entities, and (b) over timestamps. (c) Year occurrence distribution considering the years 1950 to 2050 with a logarithmic scaling on the y-axis.

An example can be constructed with the help of Figure 1: Establishing a timeline using the first criterion and requiring entities e_1 and e_2 to be in the networks, yields the left (green) and right (blue) graph as a result, while the middle (red) graph is discarded, since e_2 is not contained in the graph. Utilizing the second criterion, and demanding that an edge exists between e_1 and e_2, yields only the left graph, since it is the only one in which both entities occur and are directly connected. By utilizing entity-centric timelines, the focus is laid on certain entities, or on relationships between entities. The first criterion creates a timeline that includes only those points in time the entity co-occurred with; the second one a timeline that shows points in time where two (or more) entities occurred, and where they possibly interacted with each other. With the help of the function **sent** these interactions can be analyzed further, since it is possible do display all textual co-occurrences of two entities around a certain date in the document collection.

3.2.2 Zooming

For zooming, the partitioning of the dates D into different granularities is utilized. Since there exists a distinct hierarchy for these dates, time-centric graphs for timestamps of finer granularities are necessarily subgraphs of all time-centric graphs of coarser temporal resolution (if edge and node weightings are disregarded). For zooming, the user can start from an arbitrary network G, identifying relevant relationships and entities. Zooming can then be divided into the two scenarios of zooming in and out: On the one hand, by employing a zooming out operation, the respective coarser network is displayed, which is a supergraph of G. In this supergraph, the respective subgraph G can be highlighted, but also a broader context can be explored by observing how certain relationships are embedded in the bigger picture. On the other hand, by zooming in, and given the same network G, the user can select a network of finer granularity that ranges in the same temporal interval as G. For example, given the graph associated with *June 2000*, the user can select one of the days from *June 2000* for which a graph exists, investigating the origin of specific relations, and being able to identify crucial parts of the documents by utilizing the function **sent**.

4 Experimental Results

In this section, we describe the data set used for evaluation and present the results to demonstrate the usefulness of the approach.

4.1 Description of Data Set

For evaluation, we utilize a German document collection containing juridical protocols of the NSU (*National Socialist Underground*, or *Nationalsozialistischer Untergrund* in German) trial extracted from NSU Watch[1], which also gives an introduction to the case. The NSU data set covers 387 documents, each representing one of the 437 trial days, and consisting of 180,887 sentences and 974,892 words. Protocols for fifty trial days are missing, because they are not available from NSU Watch. Most of the omitted documents are detailing the last 100 days of the trial. Additionally, we preprocess the documents by removing stop words and out-of-vocabulary tokens.

[1]https://www.nsu-watch.info/2013/05/sitzungstermine/; accessed 3. January 2020; The used data set is actually a cleaned version of the extracted data, and all results presented here are in regard to this cleaned version.

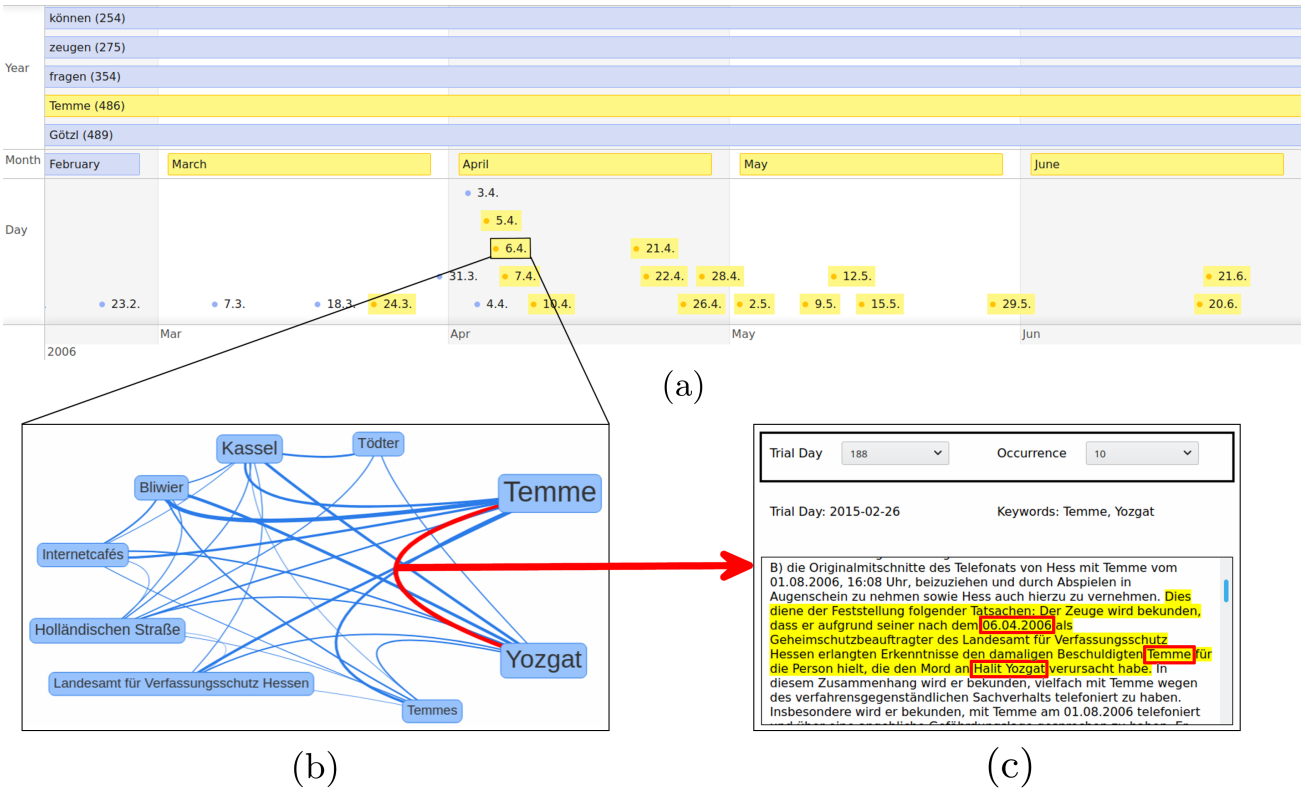


Figure 3: Illustration of a typical exploration process. (a) Excerpt of the constructed timeline, showing different time granularities. By selecting one of the five most frequently occurring words in a year all networks are marked that contain that term. (b) Clicking on April 6 shows the associated time-centric graph reduced to the 10 nodes with the highest tf-itf score, ignoring date nodes. Size of nodes and edges depends on assigned weights. (c) Clicking on an edge allows the user to browse term co-occurrences in the vicinity of temporal tags (red highlights) in documents.

Timestamp generation is done by temporal tagging employing HeidelTime [SG13], resulting in total in 15,104 date instances. After the extraction of time-centric co-occurrence graphs around these individual dates, utilizing a window size of 4 sentences, our method yields a total of 1072 networks, 859 having day granularity, 191 month granularity, and 22 year granularity. Further text processing, e.g., sentence splitting or named entity recognition, is done with the help of spaCy [HM], employing the *de_core_news_md* model. For evaluation purposes, we remove the node of the associated date from each graph, since its co-occurrence with all terms in the network is trivial. Figure 2 depicts the occurrence distribution over mostly person and location entities and timestamps in the data set as well as a year occurrence distribution, showing that most of the extracted timestamps range between 1990 and 2020, which coincides with the period of time most relevant to the activities of the NSU and the trial. It can also be observed that a few dates lie in the future, which is mostly due to errors in HeidelTime's tagging.

4.2 Timeline Exploration

The focus of this work lies on the exploration of document collections, hence, we present a typical scenario of how our model is applied. Figure 3 illustrates a timeline constructed using the data described above. By searching for certain keywords one can highlight time periods, or points in time, the keyword is part of, and thus limit a search to networks one is interested in. These networks can then be analyzed manually, potentially identifying other entities relevant to the topic under investigation, or finding relations associated with a certain date. These relations can be further examined utilizing the function **sent**, such that the co-occurrences of two terms in the data set are illustrated and highlighted. Note that the networks also serves an index to the documents and sentences in which (co-occurring) dates and terms occur.

Table 1: The dates, victims and cities associated with the murders of the NSU as well as the total number of occurrences of the entity in the document collection. Only Yozgat is among the 100 most-occurring terms in the text. The tf-itf score always refers to the rank of the term in the associated time-centric co-occurrence graph.

Date	Victim	#Occs	tf-itf rank	City	#Occs	tf-itf rank
September 9, 2000	Şimşek	132	5	Nuremberg	239	6
June 13, 2001	Özüdoğru	91	1	Nuremberg	239	2
June 27, 2001	Taşköprü	79	1	Hamburg	88	8
August 29, 2001	Kılıç	112	1	Munich	213	4
February 25, 2004	Turgut	84	2	Rostock	81	1
June 9, 2005	Yaşar	107	1	Nuremberg	239	2
June 15, 2005	Boulgarides	82	1	Munich	213	2
April 4, 2006	Kubaşık	165	1	Dortmund	218	2
April 6, 2006	Yozgat	395	2	Kassel	291	3
April 25, 2007	Kiesewetter	143	4	Heilbronn	149	1

Table 2: The three highest tf-itf ranks for the two cities (a) Kassel, and (b) Nuremberg.

Date	tf-itf value
April 6, 2006	0.000806
March 18, 2006	0.000303
April 4, 2006	0.000120

(a)

Date	tf-itf value
June 13, 2001	0.000287
June 9, 2005	0.000221
September 9, 2000	0.000175

(b)

4.3 Day-centric Evaluation

For evaluation purposes, we investigate the results for the tf-itf ranking for certain key events of the NSU crimes. Table 1 gives an overview of the victims and places of the 10 murders committed by the NSU, also stating the number of occurrences and the respective tf-itf rank in the associated time-centric network. It should be expected that such key persons and locations are well represented in the constructed time-centric graphs. And indeed, one can observe that for all murder dates, the name of the victim has at least tf-itf rank 5, most of them even rank first. While not as predominant as the name of the victims, the respective locations of the murders also rank very high in regard to their tf-itf scores. Hence, one can expect that the constructed time-centric co-occurrence graphs adequately represent the events discussed during the trial. Table 2 shows the dates for which the two cities Kassel and Nuremberg have the highest tf-itf scores. Comparison with Table 1 indicates that the three major dates for Nuremberg are all associated with a murder during the respective day. For Kassel, the by far most prominent date is the date of the murder of Halit Yozgat. The two other dates are shortly before the incident, with March 18, 2006, being the day of a right-wing extremist concert discussed during the trial.

5 Conclusion and Ongoing Work

In this paper, we introduced time-centric co-occurrence networks, and presented a framework based on these networks that enables users to explore document collections using a timeline. We also introduced two applications of the proposed model, entity-centric timelines and zooming operations. The method was then applied to a collection of court protocols of the NSU trial, and we demonstrated the usefulness of our approach by showing that persons and cities relevant to the trial are well represented in our model. As future work, we aim to refine the employed edge weighting technique, e.g., taking into account the distance between two words when extracting co-occurrences, and hence, extending the possibilities for more complex analyses of entity relationships.

References

[ABYG07] Omar Alonso, Ricardo Baeza-Yates, and Michael Gertz. Exploratory Search using Timelines. In *SIGCHI 2007 Workshop on Exploratory Search and HCI Workshop*, volume 1, pages 1–4, 2007.

[ASBYG11] Omar Alonso, Jannik Strötgen, Ricardo A Baeza-Yates, and Michael Gertz. Temporal Information Retrieval: Challenges and Opportunities. *Temporal Web Analytics Workshop*, 11:1–8, 2011.

[CDJJ14] Ricardo Campos, Gaël Dias, Alípio M Jorge, and Adam Jatowt. Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)*, 47(2):1–41, 2014.

[HM] Matthew Honnibal and Ines Montani. Spacy: Industrial-Strength Natural Language Processing, version 2.1.8, https://spacy.io/, accessed 17. March 2020.

[KMN98] B Knight, J Ma, and E Nissan. Representing Temporal Knowledge in Legal Discourse. *Information and Communications Technology Law*, 7(3):199–211, 1998.

[LSCO10] Nikolaos Lagos, Frederique Segond, Stefania Castellani, and Jacki O'Neill. Event Extraction for Legal Case Building and Reasoning. In *International Conference on Intelligent Information Processing*, pages 92–101. Springer, 2010.

[PSW12] Natalia Prytkova, Marc Spaniol, and Gerhard Weikum. Predicting the evolution of taxonomy restructuring in collective web catalogues. In *WebDB*, pages 49–54, 2012.

[SG13] Jannik Strötgen and Michael Gertz. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2):269–298, 2013.

[SG15] Jannik Strötgen and Michael Gertz. A Baseline Temporal Tagger for all Languages. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 541–547, 2015.

[SM19] Julius Steen and Katja Markert. Abstractive Timeline Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 21–31, 2019.

[SSBG15] Andreas Spitz, Jannik Strötgen, Thomas Bögel, and Michael Gertz. Terms in Time and Times in Context: A Graph-based Term-Time Ranking Model. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Companion Volume*, pages 1375–1380, 2015.

[Wik20a] Wikipedia contributors. Cumex-files — Wikipedia, the free encyclopedia, 2020. [Online; accessed 26-January-2020].

[Wik20b] Wikipedia contributors. Enron scandal — Wikipedia, the free encyclopedia, 2020. [Online; accessed 26-January-2020].

[Wik20c] Wikipedia contributors. National socialist underground — Wikipedia, the free encyclopedia, 2020. [Online; accessed 26-January-2020].

[Wik20d] Wikipedia contributors. Panama papers — Wikipedia, the free encyclopedia, 2020. [Online; accessed 26-January-2020].