Report on the 12th Temporal Web Analytics Workshop (TempWeb 2022) at WWW 2022

Marc Spaniol
University of Caen Normandy
Caen, France
marc.spaniol@unicaen.fr

Ricardo Baeza-Yates
Northeastern University
Silicon Valley, CA, USA
rbaeza@acm.org

Omar Alonso
Amazon
Santa Clara, CA, USA
oralonso@gmail.com

Abstract

TempWeb focuses on investigating infrastructures, scalable methods, and innovative software for aggregating, querying, and analyzing heterogeneous data at Web scale. Emphasis is given to data analysis along the time dimension for web data that has been collected over extended time periods. A major challenge in this regard is the sheer size of the data it exposes and the ability to make sense of it in a useful and meaningful manner for its users. It is worth noting that this trend of using big data to make inferences is not specific to Web content analytics, so work presented here might be useful in other areas, too. TempWeb has been created for this purpose and is the ideal venue to discuss all its facets.

Date: 25 May, 2022.

Website: http://temporalweb.net/.

1 Introduction

TempWeb 2022 was the twelfth event in its workshop series and took place co-located on 25th May 2022 in conjunction with The Web Conference WWW 2022[1]. As in previous years, they were intended to be held as a potentially hybrid event, this time in Lyon, but due to the pandemic, were fully moved online.

The focus of TempWeb and the topics addressed are a "natural" match with The Web Conference. With digital content born almost two decades ago, the need for a more systematic exploitation of our digital cultural heritage as well as new analysis techniques, becomes evident. Hence, TempWeb is an ideal venue to exchange knowledge about temporal analytics at a Web scale with experts from science and industry. Further workshop details can be found on the workshop homepage[2]. The program of TempWeb 2022 was organized as a full day workshop in four sessions:

Session 1 Workshop Introduction and Keynote by Adam Jatowt (University of Innsbruck)

Chair: Marc Spaniol (University of Caen Normandy, France)

[1] https://www2022.thewebconf.org/
[2] http://temporalweb.net/

Session 2 Time in News and on the Web
Chair: Lirong Zhang (University of Tsukuba, Japan)
Session 1 Temporal Networks and Models
Chair: Adam Jatowt (University of Innsbruck)
Session 1 Panel with Workshop Organizers and Keynote Speaker
Chair: Ricardo Baeza-Yates (Northeastern University, Silicon Valley, CA, USA)
Information about the detailed program of the workshop is available at the conference website[3].

2 Workshop Objectives and Topics

The objective of TempWeb is to provide a venue for researchers of all domains (IE/IR, Web mining, etc.) where the temporal dimension opens an entirely new range of possibilities and challenges. The workshop's ambition is to keep shaping a community of interest on research issues resulting from the introduction of the time dimension in web mining and analysis. The maturity of the Web as well as the emergence of large-scale repositories of web content makes this very timely and as a result, a growing number of research projects and services are emerging that have this focus in common. Having a dedicated workshop will help, we believe, to take a rich and cross-domain approach to this continuous research with a strong focus on the temporal dimension.

To this end, TempWeb focuses on investigating infrastructures, scalable methods, and innovative software for aggregating, querying, and analyzing heterogeneous data at Web scale. Emphasis will be given to temporal data analysis along the time dimension for web data that has been collected over extended time periods. A major challenge in this regard is the sheer size of the data it exposes and the ability to make sense of it in a useful and meaningful manner for its users. As such, longitudinal aspects in Web content analysis become relevant for analysts from various domains, including, but not limited to sociology, marketing, environmental studies, politics, etc. Studies in this context range from "low-level" structural network log analysis over time, up to "high-level" entity-level Web content analytics and terminology evolution. While both before mentioned aspects represent the extremes of the spectrum, they have one thing in common: Web scale data analytics needs to develop infrastructures and extended analytical tools in order to make use of that data. To this end, workshop topics of TempWeb therefore include, but are not limited to following:

- Web scale data analytics
- Temporal Web analytics
- Distributed data analytics
- Web science
- Web dynamics
- Data quality metrics
- Web spam evolution
- Content evolution on the Web
- Systematic exploitation of Web archives
- Large scale data storage

[3] https://www2022.thewebconf.org/conference-schedule/

- Large scale data processing
- Time aware Web archiving
- Data aggregation
- Web trends
- Topic mining
- Terminology evolution
- Community detection and evolution

3 Workshop Contributions

For its twelfth edition, TempWeb accepted six very positively reviewed submissions for oral presentation (acceptance rate of 75%). We interpret the high quality of the submissions and the frequent contributors to TempWeb, as indicators of an evolving community. It shows a clear sign of a positive dynamic in the study of time in the scope of the Web and evidence of the relevance of this effort.

This edition's keynote talk on *Temporal Question Answering in News Article Collections* was given by Prof. Adam Jatowt (University of Innsbruck) [Jatowt, 2022]. In his talk he addressed the challenges of open-domain question answering when answering user questions against a large document collection. He highlighted that current approaches usually utilize corpora of relatively short time-spans. To this end, he introduced a novel approach overcoming this limitation by incorporating temporal aspects inherent in the corpus and queries. Therefore, he presented an approach that combines temporal information retrieval with natural language processing. Further, the relevance of accurate dating was discussed. The talk was concluded by the presentation of a novel large-scale question answering dataset called ArchivalQA.

In their paper on *A Bi-level assessment of Twitter data for election prediction: Delhi Assembly Elections 2020* Maneet Singh, S.R.S. Iyengar, Akrati Saxena and Rishemjit Kaur presented a recent study utilizing Twitter to assess the outcome of the Delhi Assembly elections 2020 [Singh et al., 2022]. To this end, the investigated election results with the activities of different candidates and parties on Twitter by incorporating mentions and sentiment of voters' responses along the temporal dimension. This resulted in the observation that considering the number of followers and the replies to candidates' tweets proved to be good indicators for predicting actual in order to predict the result. At the same time, counts of party mentioning and the temporal analysis of voters' sentiment towards the party was not aligned with the election result.

Oded Ovadia, Oren Elisha and Elad Yom-Tov presented the *Detection of Infectious Disease Outbreaks in Search Engine Time Series Using Non-Specific Syndromic Surveillance with Effect-Size Filtering* [Ovadia et al., 2022]. Therefore, they studied how non-specific syndromic surveillance systems might be used in order to detect novel infectious disease outbreaks, such as COVID-19. Therefore, the main idea is to employ and aggregate data from various data sources such as official electronic health records or even logged search engine queries. Obviously, this has particular relevance when symptoms of a novel disease are vague.

The question *Why Round Years are Special? Analyzing Time References in News Article Collections* was presented by Adam Jatowt, Antoine Doucet and Ricardo Campos [Jatowt et al., 2022]. To this end, they studied a news article collection spanning 34 years by investigating

time expressions and their interplay with the corresponding publication dates. In particular, they employed the co-occurring named entities as representation over a graph-based representation of temporal expressions. As a result, they observed five key findings that could be incorporated in various applications employing temporal expressions.

A *Multi-touch Attribution for complex B2B customer journeys using Temporal Convolutional Networks* was presented by Aniket Agrawal, Nikhil Sheoran, Sourav Suman and Gaurav Sinha [Agrawal et al., 2022]. To this end, the authors' introduced a deep learning-based framework that aims at resolving attribution problems through modeling conversions of journeys as functions of stage transitions. In order to do so, the model connects a temporal convolutional network (TCN) with a global conversion model stage-TCN. The conducted experiments on two real-world B2B datasets show the viability of the proposed approach. Further, perturbation-based techniques highlight the robustness of the presented approach.

The paper by Lirong Zhang, Hideo Joho and Hai-Tao Yu introduced the *Semantic Modelling of Document Focus-time for Temporal Information Retrieval* [Zhang et al., 2022]. In particular, they aimed at a better understanding of the factor time in order to improve retrieval. To this end, they introduced a novel method to estimate the focus-time of documents leveraging their semantic information as well as to understand the temporal intend utilizing Google Trend. As a proof of concept, they conducted a study on temporal information retrieval and temporal diversity retrieval. The results on the NTCIR Temporalia data set show the viability of their approach and its effectiveness.

Last, but not least, an *Analytical Models for Motifs in Temporal Networks* by Alexandra Porter, Baharan Mirzasoleiman and Jure Leskovec was presented [Porter et al., 2022]. The paper is motivated by the massive computational costs of identifying temporal motifs in domains such as social networks, communication networks, and financial transaction networks. Therefore, they introduce a fast and accurate model-based method for counting motifs in temporal networks. To this end, they combine a temporal activity state block model in order to derive closed-form analytical expressions. Thus, motif frequencies and variances can be rapidly computed. The viability of their approach has then been validated through experiments on two large real-world networks.

4 Conclusion and Future Directions

As in previous years, the TempWeb 2022 was highly interactive. Many discussions emerged from the presented papers and were brought over to the panel. Not surprisingly, the many facets of temporal Web analytics became evident and showed the importance of their investigation at a venue like TempWeb. The vast majority of participants regretted the opportunity to meet face to face and having the opportunity to organize a social event to foster networking.

In the meanwhile, the proposal for TempWeb 2023 has been accepted for the upcoming edition of The Web Conference. Thus, TempWeb 2023 is scheduled to take place as an in presence event on May 1/2 in conjunction with The Web Conference 2023 (WWW 2023) in Austin, TX, USA.

Acknowledgements

TempWeb thanks the authors of submitted papers for their efforts, and all workshop attendees. Further, we thank our session chairs for helping in securing a smooth flow of the Workshop. In addition, we would like to express our gratitude to Adam Jatowt (University of Innsbruck, Austria) for providing you with an inspiring keynote talk and his participation in the concluding panel.

TempWeb 2022 was jointly organized by University of Caen Normandy (Caen, France) and Northeastern University (Silicon Valley, CA, USA). We also wish to express our particular thanks to our program committee members:

- Andras A. Benczur (Hungarian Academy of Science)
- Behrooz Mansouri (Rochester Institute of Technology, USA)
- Klaus Berberich (University of Applied Sciences Saarbrücken, Germany)
- Ricardo Campos (Polytechnic Institute of Tomar)
- Govind (Amazon, India)
- Adam Jatowt (University of Innsbruck, Austria)
- Nattiya Kanhabua (Upwork, Bangkok, Thailand)
- Scott Kirkpatrick (Hebrew University Jerusalem, Israel)
- Amit Kumar (Université de Caen Normandy, France)
- Frank McCown (Harding University, USA)
- Michael Nelson (Old Dominion University, USA)
- Nikos Ntarmos (Huawei Technologies R&D, UK)
- Kjetil Nørvåg (Norwegian University of Science and Technology, Norway)
- Thomas Risse (University Library Johann Christian Senckenberg, Germany)
- Andreas Spitz (University of Konstanz, Germany)
- Jannik Strötgen (Bosch Center for Artificial Intelligence, Germany)
- Torsten Suel (NYU Polytechnic, USA)
- Masashi Toyoda (Tokyo University, Japan)
- Gerhard Weikum (Max-Planck-Institut für Informatik, Germany)

Last but not least, we would like to thank the organizers of WWW2022 for hosting the workshops and providing the online meeting facilities.

References

Aniket Agrawal, Nikhil Sheoran, Sourav Suman, and Gaurav Sinha. Multi-touch attribution for complex B2B customer journeys using temporal convolutional networks. In Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini, editors, *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*, pages 910–917. ACM, 2022. doi: 10.1145/3487553.3524670. URL https://doi.org/10.1145/3487553.3524670.

Adam Jatowt. Temporal question answering in news article collections. In Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini, editors, *Companion of The Web Conference 2022, Virtual Event / Lyon, France,*

April 25 - 29, 2022, page 895. ACM, 2022. doi: 10.1145/3487553.3526023. URL https://doi.org/10.1145/3487553.3526023.

Adam Jatowt, Antoine Doucet, and Ricardo Campos. Diachronic analysis of time references in news articles. In Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini, editors, *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*, pages 918–923. ACM, 2022. doi: 10.1145/3487553.3524671. URL https://doi.org/10.1145/3487553.3524671.

Oded Ovadia, Oren Elisha, and Elad Yom-Tov. Detection of infectious disease outbreaks in search engine time series using non-specific syndromic surveillance with effect-size filtering. In Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini, editors, *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*, pages 924–929. ACM, 2022. doi: 10.1145/3487553.3524672. URL https://doi.org/10.1145/3487553.3524672.

Alexandra M. Porter, Baharan Mirzasoleiman, and Jure Leskovec. Analytical models for motifs in temporal networks. In Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini, editors, *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*, pages 903–909. ACM, 2022. doi: 10.1145/3487553.3524669. URL https://doi.org/10.1145/3487553.3524669.

Maneet Singh, S. R. S. Iyengar, Akrati Saxena, and Rishemjit Kaur. A bi-level assessment of twitter data for election prediction: Delhi assembly elections 2020. In Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini, editors, *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*, pages 930–935. ACM, 2022. doi: 10.1145/3487553.3524673. URL https://doi.org/10.1145/3487553.3524673.

Lirong Zhang, Hideo Joho, and Hai-Tao Yu. Semantic modelling of document focus-time for temporal information retrieval. In Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini, editors, *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*, pages 896–902. ACM, 2022. doi: 10.1145/3487553.3524668. URL https://doi.org/10.1145/3487553.3524668.