Report on the 13th Workshop on Temporal Web Analytics (TempWeb 2023) at WWW 2023

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

Ricardo Baeza-Yates  
Northeastern University  
USA  
rbaeza@acm.org

Omar Alonso  
Amazon  
USA  
oralonso@gmail.com

Abstract

TempWeb is an established Workshop (series) with a long-standing tradition as a co-located event at The Web Conference. Considering the constantly evolving Web as a primary object of research, TempWeb provides a platform to present longitudinal studies on the Web (structure), its contents, and its communities, etc. by aiming at the investigation of infrastructures, scalable methods, and innovative software for aggregating, querying, and analyzing heterogeneous data at Web scale. Since longitudinal aspects in Web analytics is becoming more and more inter- and transdisciplinary relevant for analysts from various domains, the studies, tools, and demonstrations are not only limited to computer science, but also open to sociology, marketing, environmental studies, politics, etc., to name just a few. As such, TempWeb has developed as a forum for a community from academia and industry covering temporal Web analytics at all levels of granularity and across various disciplines.

Date: 1 May 2023.
Website: [http://temporalweb.net/](http://temporalweb.net/).

1 Introduction

TempWeb 2023 was the thirteenth event in its workshop series and took place co-located on 1st May 2023 in conjunction with The Web Conference WWW 2023. After three tough years in “quarantine”, TempWeb 2023 and WWW 2023 were the first editions in a post pandemic era and intended to be held primarily as onsite events, this time in Austin, TX, USA. However, due to organizational constraints both events took place in a partial hybrid mode with an emphasis on onsite participation.

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.

\[1\text{https://www2023.thewebconf.org/}\]
\[2\text{http://temporalweb.net/}\]
The program of TempWeb 2023 was organized as a full day workshop in five sessions:

- Session 1: Workshop Welcome and Introduction
- Session 2: Temporal Social Media
- Session 3: Communities & Time
- Session 4: Events and the Temporal Dimension
- Session 5: Open Discussion and Final Remarks

Further information about the program of the workshop is available at the conference website.

2 Workshop Objectives and Topics

The objective of TempWeb was 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 was 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 made this very timely and as a result, a growing number of research projects and services were emerging that have this focus in common. Having a dedicated workshop helped to take a rich and cross-domain approach to this continuous research with a strong focus on the temporal dimension.

TempWeb focused on investigating infrastructures, scalable methods, and innovative software for aggregating, querying, and analyzing heterogeneous data at Web scale. Emphasis was given to temporal data analysis along the time dimension for Web data that had been collected over extended time periods. A major challenge in this regard was the sheer size of the data it exposed 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 became relevant for analysts from various domains, including, but not limited to sociology, marketing, environmental studies, politics, etc. Studies in this context ranged 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 needed and still needs to develop infrastructures and extended analytical tools in order to make use of that data. To this end, workshop topics of TempWeb therefore included, but were not limited to the 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
- Large scale data processing
- Time aware Web archiving
- Data aggregation
- Web trends
- Topic mining
- Terminology evolution
- Community detection and evolution

3https://archives.iw3c2.org/www2023/program/detailed-program/
3 Workshop Contributions

For its thirteenth edition, TempWeb accepted eight very positively reviewed submissions for oral presentation (acceptance rate of \(\sim 70\%\)) [Spaniol et al., 2023a]. The high quality of the submissions and the frequent contributors to TempWeb can be interpreted, 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 was started with a Welcome and Introduction session by the Workshop organizers. The statements were intended to outline the workshop objectives and to introduce the agenda for the upcoming day. The workshop was decomposed into three technical sessions and an open discussion session concluding the workshop. In order to give a brief overview on the contents presented at TempWeb 2023, we summarize the key contributions of the accepted papers in the chronological order of their presentation [Spaniol et al., 2023b].

First, Farnoosh Hashemi, Ali Behrouz and Milad Rezaei Hajidehi presented CS-TGN: Community Search via Temporal Graph Neural Networks [Hashemi et al., 2023]. Particularly, they investigated community search through identification of densely connected subgraph containing query nodes. To this end, they introduced a query-driven temporal graph convolutional network that overcomes shortcomings of adopting pre-defined subgraph patterns and solely incorporating the aggregation of disjoint structural information for quality assessment. As a result, CS-TGN is capable of identifying community structures by intertwining local structures and global graph snapshots at each time point in combination with contextual attention in order to update the embeddings over time.

Next, Ke Li, Zhiwen Yu, Ying Zhang and Bin Guo addressed the Impact of COVID-19 Pandemic on Cultural Products Interests [Li et al., 2023]. In contrast to user-driven surveys, their approach aims at exploiting user-oriented data on the Web. To this end, they conducted a study over search interests in various countries and across a multitude of domains. The results highlight a strong correlation in interests and mobility reduction. Particularly, an increase of dematerialized media was recognizable.

The societal thread of influence campaigns was then highlighted in Detecting the Hidden Dynamics of State-Backed Influencers Using Temporal Correlations by Keeley Erhardt and Dina Albassam [Erhardt and Albassam, 2023]. While existing approaches are mostly constrained to target orchestrated campaigns, the authors presented a novel generative model that allows to reveal coordinated inauthentic activity of individual social media accounts. To this end, their model describes the interactions between networked Markov chains.

In The Age of Snippet Programming: Towards Understanding Developers Communities in Stack Overflow and Reddit, Alessia Antelmi, Gennaro Cordasco, Daniele De Vinco and Carmine Spagnuolo analyze the importance of community contribution-based question-and-answer (Q&A) platforms as an information source of coding related knowledge exchange [Antelmi et al., 2023]. To this end, they conducted longitudinal study on the posting activities of users and their high-order interaction patterns abstracted via hypergraphs. Their results reveal significant differences between Stack Overflow and Reddit. In particular, they highlight a trend towards highly community-friendly platforms, e.g., Reddit.

Subsequently, Behrooz Mansouri, Ricardo Campos and Adam Jatowt introduced a novel approach Towards Timeline Generation with Abstract Meaning Representation [Mansouri et al.,
In their paper, the authors employ a graphical representation of the text where the nodes are semantic concepts and the edges denote relationships between concepts. To this end, sentences are selected that contain the highest number of named-entities and are then sorted based on a combined score of Inverse Document Frequency (IDF) of graph nodes and a keyword extraction method. Experiments on a COVID-19 data set showed the viability of the approach.

In Gone, Gone, but Not Really, and Gone, But Not Forgotten: A Typology of Website Recoverability Brenda Reyes Ayala studies the recoverability of Web pages [Ayala, 2023]. To this end, four Web archive collections were used in order to analyze the content and evolution of previously thought to be lost Web pages. This resulted in three types of lost contents: not recoverable, fully recoverable and partially recoverable. Based on the before introduced categories, Web archivists were supported in finding those contents that were previously considered to be lost yielding to a more complete Web archive.

A novel approach of reliable discovering new content from Web sites was introduced by Peter Schulam and Ion Muslea in Improving the Exploration/Exploitation Trade-Off in Web Content Discovery [Schulam and Muslea, 2023]. They observed that SOTA implementations employ probabilistic methods that result in a sub-optimal balancing of exploration and exploitation. In order to overcome this deficit they developed a new crawler called Thompson, named after the sampler that drives its refresh decisions. In experiments they were then able to show that their approach outperforms the SOA in various dimensions due to a better modeling of the action space.

Finally, Aparna Varde, Gerard de Melo and Boxiang Dong presented their research on Temporal Ordinance Mining for Event-Driven Social Media Reaction Analytics [Varde et al., 2023]. For that purpose, they proposed a research agenda enabling novel forms of temporally-driven ordinance content analysis and Web question answering. In combination with an ongoing social media analysis this will ultimately enable historical ordinance mining. As a result, the overall process will become more transparent for both legislators and the broader public.

4 Conclusion and Future Directions

As in previous years, the TempWeb 2023 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. Participants were very engaged in the discussions and asked many times for a continuation of running the workshop, which shows a continuous interest on the topic.

Based on the very positive feedback, a proposal for TempWeb 2024 was submitted and later accepted for the upcoming edition of The Web Conference. Thus, TempWeb 2024 will take place as an in presence event to be held on May 14 in Singapore. In addition, a special issue on “The Temporal Web: Investigating Time and the Temporal Dimension” has been released at ACM Transactions on The Web (cf. the CfP for details).

Acknowledgements

TempWeb thanks the authors of submitted papers for their efforts, and all workshop attendees. TempWeb 2023 was jointly organized by University of Caen Normandy (Caen, France), Northeastern University (Silicon Valley, CA, USA) and Amazon, Santa Clara, CA, USA. We also wish to express our particular thanks to our program committee members:

- Eytan Adar (University of Michigan, USA)
- Srikanta Bedathur (IIT Delhi, India)
- Andras A. Benczur (Hungarian Academy of Science)
- Behrooz Mansouri (Rochester Institute of Technology, USA)
- 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 (Tarento Technologies Pvt Limited, India)
- 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 WWW2023 for hosting the workshops and providing the online meeting facilities.

References


Brenda Reyes Ayala. Gone, Gone, but Not Really, and Gone, But Not forgotten: A Typology of Website Recoverability. In Ying Ding, Jie Tang, Juan F. Sequeda, Lora Aroyo, Carlos Castillo, and Geert-Jan Houben, editors, Companion Proceedings of the ACM Web Conference 2023,


