The 4th Temporal Web Analytics Workshop (TempWeb'14)

Marc Spaniol
Max-Planck-Institut für Informatik
Saarbrücken
Germany
mspaniol@mpi-inf.mpg.de

Julien Masanès
Internet Memory Foundation
Paris
France
julien@internetmemory.org

Ricardo Baeza-Yates
Yahoo Labs
Barcelona
Spain
rbaeza@acm.org

ABSTRACT
In this paper we give an overview on the 4th Temporal Web Analytics Workshop (TempWeb). The goal of TempWeb is to provide a venue for researchers of all domains (IE/IR, Web mining, etc.) where the temporal dimension opens up an entirely new range of challenges and possibilities. The workshop's ambition is to help shaping a community of interest on the research challenges and possibilities resulting from the introduction of the time dimension in web analysis. Having a dedicated workshop will help, we believe, to take a rich and cross-domain approach to this new research challenge with a strong focus on the temporal dimension. For the fourth time, TempWeb has been organized in conjunction with the International World Wide Web (WWW) conference, being held on April 8, 2014 in Seoul, Korea.

Categories and Subject Descriptors
H.3.1: Content Analysis and Indexing

General Terms
Algorithms, Management, Measurement, Documentation, Experimentation

Keywords
Temporal Web Analytics, Web Scale Data Analytics, Distributed Data Analytics

1. INTRODUCTION
After three successful editions, the TempWeb workshop's specific focus on temporal dimension is getting more and more relevant. Established fields of research (IE/IR, Web mining, etc.) are challenged to leverage time signals and expressions to capture dynamics and trends and include contextualized time. The maturity of the Web, the emergence of large scale repositories of web material, makes this very timely and a growing set of research and services are emerging that have this focus in common. Having a dedicated workshop has proven relevant and fruitful to take a rich and cross-domain approach to this new research challenge that focus on the temporal dimension.

The focus of TempWeb and the topics addressed are a "natural" match with the WWW conference. With digital content born almost two decades ago, the need for a more systematic exploitation of our digital cultural heritage becomes evident. While the early 90's of the Web have been almost completely lost, national libraries, digital news archives and archiving institutions (like the Internet Memory Foundation) have protected Web contents from vanishing. These data are a potential goldmine for temporal and large-scale Web analytics at the content level. However, the societal as well as scientific impacts of temporal Web analytics have been not sufficiently studied. As the WWW conference is the premier event series in this domain, we consider TempWeb an ideal venue to exchange knowledge about temporal analytics at Web scale with experts from science and industry.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579047

2. WORKSHOP TOPIC AND THEMES
TempWeb focuses on investigating infrastructures, scalable methods, and innovative software for aggregating, querying, and analyzing heterogeneous data at web scale. Particular emphasis is 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. Web scale data analytics therefore needs to develop infrastructures and extended analytical tools to make sense of the mass of information that the historic and current web represents. Topics of TempWeb therefore include, but are not limited to the following:

- Web scale data analytics
- Temporal Web analytics
- Distributed data analytics
- 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

3. WORKSHOP STRUCTURE

This year, six high quality submissions were finally accepted for oral presentation. Apart from a keynote talk by Masashi Toyoda (Tokyo University, Japan) and a panel discussion concluding the workshop, the contributions covered the full spectrum of temporal Web analytics ranging from data provisioning, up to higher-level semantics in order to making sense of temporal Web data.

Three contributions addressed aspects related to "Data Provisioning". Jimmy Lin, Milad Gholami and Jinfeng Rao presented an "Infrastructure for Supporting Exploration and Discovery in Web Archives". To this end, they developed a tool called Warcbase, which is an open-source platform for managing Web archives built on the distributed datastore HBase. With their system and the underlying data model it becomes possible to store and manage raw content as well as metadata and extracted knowledge. The paper on "NTCIR Temporalia: A Test Collection for Temporal Information Access Research" by Hideo Joho, Adam Jatowt and Roi Blanco introduced a novel standardized evaluation benchmark for fostering research in Temporal IR. For that reason, they developed Temporal Information Access (Temporalia), which is a new pilot task run at NTCIR-11. On the ontological-level, Gaël Dias, Mohammed Hasanuzzaman, Stéphane Ferrari and Yann Mathet introduced "TempoWordNet for Sentence Time Tagging". TempoWordNet is an approach towards a temporal ontology, which may contribute to the success of time-related applications. The underlying idea is to learn temporal classifiers from a set of time-sensitive synsets and then to apply them to the entire WordNet.

The remaining three papers of the workshop aimed at "Making Sense of Temporal Web Data". Lars Döhling and Ulf Leser introduced a novel approach toward "Extracting and Aggregating Temporal Events from Text". Their paper describes a three-step framework, which is able to extract and condense specific facts about events. They evaluated their approach in a comprehensive case study by gathering data on particular earthquakes from Web data sources. In order to help "Understanding Time through Wikipedia", Stewart Whiting and Omar Alonso gave an overview of the past work and characteristics that support Wikipedia for time-aware research. In addition, they explained the main content and meta-data temporal signals by also briefly discussing the source and nature of each signal. Last but not least, Sushma Bannur and Omar Alonso presented research on "Analyzing Temporal Characteristics of Check-in data". To this end, they conducted a large study using check-in data from Facebook to analyze different temporal characteristics in four venue categories. Based on their findings they outlined new search scenarios, where the combination of location and temporal-aware data might be beneficial.

4. ORGANIZATION

Covering this novel and challenging research area of temporal Web analytics, the workshop organizers teamed up from an archiving institution as well as industrial and academic research. Similarly, the international program committee was composed of well renowned experts in one or more of the topics addressed. The program committee consisted of the following members:

- Eytan Adar (University of Michigan, USA)
- Omar Alonso (Microsoft Bing, USA)
- Ralitsa Angelova (Google, Switzerland)
- Srikanta Bedathur (IIIT-Delhi, India)
- Andras A. Benczur (Hungarian Academy of Science)
- Klaus Berberich (Max Planck Institute for Informatics, Germany)
- Roi Blanco (Yahoo Labs, Spain)
- Philipp Cimiano (University of Bielefeld, Germany)
- Renata Galante (Universidade Federal do Rio Grande do Sul, Brazil)
- Adam Jatowt (Kyoto University, Japan)
- Scott Kirkpatrick (Hebrew University Jerusalem, Israel)
- Frank McCown (Harding University, USA)
- Michael Nelson (Old Dominion University, USA)
- Kjetil Nørvåg (Norwegian University of Science and Technology, Norway)
- Nikos Ntarmos (University of Patras, Greece)
- Philippe Rigaux (CNAM and Mignify, France)
- Thomas Risse (L3S Research Center, Germany)
- Pierre Senellart (Telecom ParisTech, France)
- Masashi Toyoda (Tokyo University, Japan)
- Peter Triantafillou (University of Glasgow, UK)
- Gerhard Weikum (Max Planck Institute for Informatics, Germany)