TempWeb 2012

Proceedings of the 2nd Temporal Web Analytics Workshop

Ricardo Baeza-Yates
Julien Masanès
Marc Spaniol

Chairs/Editors

Lyon, France
April 17, 2012

The Association for Computing Machinery
2 Penn Plaza, Suite 701
New York New York 10121-0701

ACM ISBN: 1-59593-036-1

Preface

Time is a key dimension to understand the web. It is fair to say that it has not received yet all the attention it deserves and TempWeb is an attempt to help remedy this situation by putting time as the center of its reflexion.

Studying time in this context actually covers a large spectrum, from dating methodology to extraction of temporal information and knowledge, from diachronic studies to the design of infrastructural and experimental settings enabling a proper observation of this dimension.

For its second edition, TempWeb includes 6 papers out of a total of 17 papers submitted which put it's acceptance rate at 35%. The number of papers submitted has almost doubled compared to the first edition, which we like to interpret as a clear sign of positive dynamic and an indication of the relevance of this effort. The workshop proceedings are published in ACM DL (ISBN 978-1-4503-1188-5).

We hope you will find in these papers, the keynotes and the discussion and exchanges of this edition of TempWeb some motivations to look more into this important aspect of Web studies.

TempWeb 2012 was jointly organized by Internet Memory Foundation (Paris, France), the Max-Planck-Institute for Computer Science (Saarbrücken, Germany) and Yahoo! Research (Barcelona, Spain), and supported by the 7th Framework IST programme of the European Union through the focused research project (STREP) on Longitudinal Analytics of Web Archive data (LAWA) under contract no. 258105.

Ricardo Baeza-Yates
Julien Masanès
Marc Spaniol

Lyon (France), April 2012

Chairs

Ricardo Baeza-Yates (Yahoo! Research, Spain)
Julien Masanès (internet Memory Foundation, France and Netherlands)
Marc Spaniol (Max-Planck-Institut für Informatik, Germany)

International Program Committee

Eytan Adar (University of Michigan, USA)
Omar Alonso (Microsoft Bing, USA)
Srikanta Bedathur (IIIT-Delhi, India)
András A. Benczúr (Hungarian Academy of Science)
Klaus Berberich (Max-Planck-Institut für Informatik, Germany)
Roi Blanco (Yahoo! Research, Spain)
Adam Jatowt (Kyoto University, Japan)
Scott Kirkpatrick (Hebrew University Jerusalem, Israel)
Christian König (Microsoft Research, USA)
Frank McCown (Harding University, USA)
Michael Nelson (Old Dominion University, USA)
Nikos Ntarmos (University of Patras, Greece)
Kjetil Nørvåg (Norwegian University of Science and Technology, Norway)
Philippe Rigaux (Internet Memory Foundation, France and Netherlands)
Thomas Risse (L3S Research Center, Germany)
Pierre Senellart (Télécom ParisTech, France)
Torsten Suel (NYU Polytechnic, USA)
Masashi Toyoda (Tokyo University, Japan)
Peter Triantafillou (University of Patras, Greece)
Michalis Vazirgiannis (Athens University of Economics and Business & École Polytechnique)
Gerhard Weikum (Max-Planck-Institut für Informatik, Germany)

Contents

The 2nd Temporal Web Analytics Workshop (TempWeb)

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

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

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

ABSTRACT
In this paper we give an overview on the 2nd 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 second time, TempWeb has been organized in conjunction with the International World Wide Web (WWW) conference. Hence, this year, TempWeb was held on April 17, 2012 in Lyon, France.

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms
Algorithms, Management, Measurement, Documentation, Experimentation

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

1. INTRODUCTION
With the rapidly growing amount of digitally-born contents and the Internet Archive's endeavor in capturing the World Wide Web for almost two decades, we now have more than 150 billion Web pages (> 2PB at Internet Archive - US only) at disposal. These archives not only capture the history of born-digital content but also reflect the zeitgeist of different time periods over more than a decade. This is already and will become more and more a gold mine for researchers, such as sociologists, political scientists, media and market analysts, as well as experts on industrial property (IP, e.g., at patent offices), etc.

Research on the temporal dimension of Web contents opens up great opportunities for analysts. For example, one could compare the notions of "online friends" and "social networks" as of today versus five or ten years back. Similar examples relevant for a business analyst or technology journalist could be about "tablet PC" or "online music". Similarly, the hyperlink structure of archived material can now be systematically exploited. It makes it possible to see how websites (or even domains) develop over time, whether they are affected by web spam or not, and which prevalent structures exist in general or within a certain domain.

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 Archive Foundation) have protected Web contents from vanishing. These data are a potential goldmine for temporal Web analytics at the Web scale content level. However, the societal as well as scientific impact of temporal Web analytics have been not sufficiently been 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 on the Web scale with experts from science and industry.

2. WORKSHOP TOPICS AND THEMES
TempWeb focuses on investigating infrastructures, scalable methods, and innovative software for aggregating, querying, and analyzing heterogeneous data at Internet scale. Particular 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. It is worth noting that this trend of using big data to make inferences is not specific to Web content analytics. A now-common strategy in post-genomic biology is to measure, quantitatively, the action of all (or as many as possible) of the genes at the level of the transcriptome, proteome, metabolome and phenotype, and to use computerized methods to infer gene function via various kinds of pattern recognition techniques. On the Web, we have to a large extent, also reached this point. Web scale data analytics therefore needs to develop infrastructures and extended analytical tools to make sense of these. Workshop topics of TempWeb therefore include, but are not limited to following:

- Web scale data analytics
- Temporal Web analytics
- Web scale data analytics
- Temporal Web analytics
- Distributed data analytics
- Web science
- Web dynamics

TempWeb '12, Apr 16-17 2012, Lyon, France
ACM 978-1-4503-1188-5/12/04.

- Data quality metrics
- Web spam
- Knowledge evolution on the Web
- Systematic exploitation of Web archives
- Large scale data storage
- Large scale data processing
- Data aggregation
- Web trends
- Topic mining
- Terminology evolution
- Community detection and evolution

3. WORKSHOP STRUCTURE

This year, TempWeb received 17 high quality submissions of which 6 were finally accepted for oral presentation. At an acceptance rate of 35%, the workshop was highly selective. Apart from a keynote talk by Staffan Truvé (CTO, Recorded Future) and a panel discussion concluding the workshop, the accepted papers were structured into two sessions of three papers.

The first session was focusing on "Web Dynamics". The first paper by Geerajit Rattanaritnont, Masashi Toyoda and Masaru Kitsuregawa introduced an approach toward "Analyzing Patterns of Information Cascades based on Users' Influence and Posting Behaviors". After that, Masahiro Inoue and Keishi Tajima presented research on "Noise Robust Detection of the Emergence and Spread of Topics on the Web". Finally, the third paper by Margarita Karkali, Vassilis Plachouras, Costas Stefanatos and Michalis Vazirgiannis dealt with "Keeping Keywords Fresh: A BM25 Variation for Personalized Keyword Extraction".

The second session was centered on "Identifying and Leveraging Time Information". Erdal Kuzey and Gerhard Weikum presented a novel approach toward "Extraction of Temporal Facts and Events from Wikipedia". Then, strategies for the "Identification of Top Relevant Temporal Expressions in Documents" were introduced by Jannik Strötgen, Omar Alonso and Michael Gertz. Concluding the second session, Ricardo Campos, Gaël Dias, Alípio Jorge and Célia Nunes demonstrated a study on "Enriching Temporal Query Understanding through Date Identification: How to Tag Implicit Temporal Queries?"

4. ORGANIZATION

The workshop was the second of its kind and was held again in conjunction with the international World Wide Web (WWW) conference. Covering this novel and challenging research area of temporal Web analytics, the workshop organizers teamed up from an archiving institution, industry and research. Similarly, the international program committee is composed of well renowned experts in one or more of topics addressed. The program committee consisted of the following members:

- Eytan Adar (University of Michigan, USA)
- Omar Alonso (Microsoft Bing, USA)
- Srikanta Bedathur (IIIT-Delhi, India)
- Andras Benczur (Hungarian Academy of Science)
- Klaus Berberich (Max Planck Institute for Informatics, Germany)
- Roi Blanco (Yahoo! Research, Spain)
- Adam Jatowt (Kyoto University, Japan)
- Scott Kirkpatrick (Hebrew University Jerusalem, Israel)
- Christian König (Microsoft Research, USA)
- Frank McCown (Harding University, USA)
- Michael Nelson (Old Dominion University, USA)
- Nikos Ntarmos (University of Patras, Greece)
- Kjetil Norvag (Norwegian University of Science and Technology, Norway)
- Philippe Rigaux (Internet Memory Foundation, France and Netherlands)
- Thomas Risse (L3S Research Center, Germany)
- Pierre Senellart (Télécom ParisTech, France)
- Torsten Suel (NYU Polytechnic, USA)
- Masashi Toyoda (Tokyo University, Japan)
- Peter Triantafillou (University of Patras, Greece)
- Michalis Vazirgiannis (Athens University of Economics and Business & École Polytechnique)
- Gerhard Weikum (Max Planck Institute for Informatics, Germany)

5. ACKNOWLEDGMENTS

The organization of this workshop is partially supported by the 7th Framework IST Programme of the European Union through the focused research project (STREP) on Longitudinal Analytics of Web Archive data (LAWA) under contract no. 258105 (cf. www.lawa-project.eu).