First International Workshop on Recent Trends in News Information Retrieval (NewsIR'16)

Miguel Martinez-Alvarez[1(✉)], Udo Kruschwitz[2], Gabriella Kazai[3], Frank Hopfgartner[4], David Corney[1], Ricardo Campos[5], and Dyaa Albakour[1]

[1] Signal Media, London, UK
{miguel.martinez,david.corney,dyaa.albakour}@signal.uk.com
[2] University of Essex, Colchester, UK
udo@essex.ac.uk
[3] Lumi, London, UK
gabs@lumi.do
[4] University of Glasgow, Glasgow, Scotland
frank.hopfgartner@glasgow.ac.uk
[5] LIAAD-INESC TEC, Instituto Politécnico de Tomar, Tomar, Portugal
ricardo.campos@ipt.pt
http://www.signal.uk.com, http://www.lumi.do

Abstract. The news industry has gone through seismic shifts in the past decade with digital content and social media completely redefining how people consume news. Readers check for accurate fresh news from multiple sources throughout the day using dedicated apps or social media on their smartphones and tablets. At the same time, news publishers rely more and more on social networks and citizen journalism as a frontline to breaking news. In this new era of fast-flowing instant news delivery and consumption, publishers and aggregators have to overcome a great number of challenges. These include the verification or assessment of a source's reliability; the integration of news with other sources of information; real-time processing of both news content and social streams in multiple languages, in different formats and in high volumes; deduplication; entity detection and disambiguation; automatic summarization; and news recommendation. Although Information Retrieval (IR) applied to news has been a popular research area for decades, fresh approaches are needed due to the changing type and volume of media content available and the way people consume this content. The goal of this workshop is to stimulate discussion around new and powerful uses of IR applied to news sources and the intersection of multiple IR tasks to solve real user problems. To promote research efforts in this area, we released a new dataset consisting of one million news articles to the research community and introduced a data challenge track as part of the workshop.

1 Background and Motivation

News from mainstream media outlets is often one of the most relevant, influential and powerful sources of information. This ranges from the influence that

N. Ferro et al. (Eds.): ECIR 2016, LNCS 9626, pp. 878–882, 2016.
DOI: 10.1007/978-3-319-30671-1_85

newspapers may have on elections to the reputational damage that a negative article in a well-known magazine can cause to a brand. The process of consuming news itself is constantly changing. We receive a continuous influx of news information from different sources (e.g., traditional newspapers, blogs and social media) and this has had a massive impact on the nature of information systems.

Some of the current challenges we are facing are the integration of news data with other sources of information such as social media [1]; real-time analytics [2]; processing text in multiple languages; automatic temporal summarization [3]; and scalable processing of millions of articles on a daily basis.

Following discussions at ECIR 2015 we created a forum[1] to discover if there was enough interest within the IR community for a workshop focusing on traditional media, and news data in particular. We were very happy to see that around 40 members joined the forum straightaway and that several fruitful discussions started. This was a clear indication for the strong interest in the community for organizing such a workshop. Furthermore, the discussion in the forum illustrated the diversity of topics that this workshop could explore, including:

- Traditional and social media integration
- Temporal aspects of news
- Credibility, readability and controversy
- Bias and plurality in news
- Event and anomaly detection
- Diversification
- Summarization of multiple documents
- User-generated content (e.g., using comments to enhance news retrieval)
- News recommendation
- De-duplication and clustering of news articles
- Author identification and disambiguation
- Evaluation
- Data Visualization

2 Workshop Goals

The main goal of the workshop is to bring together scientists conducting relevant research in the field of news and information retrieval. In particular, scientists can present their latest breakthroughs with an emphasis on the application of their findings to research from a wide range of areas including: information retrieval; natural language processing; journalism (including data journalism); network analysis; and machine learning. This will facilitate discussion and debate about the problems we face and the solutions we are exploring, hopefully finding common grounds and potential synergies between different approaches. We aim to have a substantial representation from industry, from small start-ups to large enterprises, to strengthen their relationships with the academic community. This also represents a unique opportunity to understand the different problems

[1] https://groups.google.com/forum/#!forum/news-ir.

and priorities of each community and to recognize areas that are not currently receiving much academic attention but are nonetheless of considerable commercial interest. Finally, to accompany the workshop, we have released a new dataset suitable for conducting research on news IR. We describe the dataset in the next section. Detailed information about the workshop can be found on the workshop website[2].

3 The Signal Media One-Million News Articles Dataset

To stimulate workshop participation (and more generally to provide a useful resource for researchers in the area), we have prepared a new dataset of one million recent news articles from a wide range of sources (The Signal Media One-Million News Articles Dataset)[3]. In contrast to many existing collections (such as Reuters-21578 and Reuters RCV1), our new dataset include news articles from a wide range of sources including global, national and local newspapers, along with magazines and blogs. This dataset is released under the standard Creative Commons license[4] to encourage re-use in diverse non-commercial research projects. Furthermore, in the call for papers, we introduced a 'data challenge track' to encourage submissions of experimental results on our new dataset. We believe that one or more shared tasks or challenges will emerge and that, with suitable refinement, these may form the basis of future workshops. Possible challenges include but are not limited to:

- detecting and summarizing events over time;
- identifying bias in news sources to different topics and/or different entities;
- identifying influencers in media coverage and visualizing information flow;
- sentiment analysis on media coverage.

4 Keynotes and Panel

We have invited two keynote speakers who can provide insights into the topic from both an industry and an academic point of view. The industry keynote speaker is **Dr. Jochen Leidner**. Jochen is currently Director of Research at Thomson Reuters, where he heads the London (UK) R&D site, which he established. He has worked in many areas including information extraction from legal, news and financial documents, search engine technology and its application to legal information retrieval, automated proofing support for contracts, sentiment analysis, rule based systems, citation analysis and social media. The academic keynote speaker is **Dr. Julio Gonzalo**. Julio is an assistant professor at UNED (Universidad Nacional de Educacin a Distancia). Julio has been recently involved in organizing the CLEF RepLab, which is an evaluation campaign for online reputation management.

[2] http://research.signalmedia.co/newsir16.
[3] http://research.signalmedia.co/newsir16/signal-dataset.html.
[4] https://creativecommons.org/.

The workshop also includes a panel discussion with members drawn from academia, from large companies and from SMEs. This includes Dr. Jochen Leidner (Thomson Reuters), Dr. Gabriella Kazai (Lumi) and Dr. Julio Gonzalo (UNED). This panel focuses on the commonalities and differences between the communities as they face related challenges in news-based information retrieval.

5 Programme Committee

The Programme Committee (PC) is formed by key researchers from industry and academia. We thank all the PC members, whose names and affiliations are listed below.

- Ramkumar Aiyengar, Bloomberg, UK
- Marco Bonzanini, Bonzanini Consulting Ltd
- Omar Alonso, Microsoft, USA
- Alejandro Bellogin Kouki, UAM, Spain
- Horatiu-Sorin Bota, University of Glasgow, UK
- Igor Brigadir, Insight Centre for Data Analytics, Ireland
- Toine Bogers, Aalborg University Copenhagen (AAU-CPH), Denmark
- Ivan Cantador, UAM, Spain
- Arjen De Vries, Centrum Wiskunde & Informatica (CWI), The Netherlands
- Ernesto Diaz Aviles, IBM Research, Ireland
- Angel Castellanos Gonzalez , UNED, Spain
- Julio Gonzalo, UNED, Spain
- David Graus, University of Amsterdam, The Netherlands
- Jon Atle Gulla, NTNU, Norway
- Charlie Hull, Flax, UK
- Alípio Jorge, University of Porto, Portugal
- Jussi Karlgren, Gavagai, Sweden
- Marijn Koolen, University of Amsterdam, The Netherlands
- David D. Lewis, David D. Lewis Consulting, USA
- Stefano Mizzaro, University of Udine, Italy
- Elaheh Momeni, University of Vienna, Austria
- Miles Osborne, Bloomberg, UK
- Filipa Peleja, Yahoo! Research, Spain
- Vassilis Plachouras, Thomson Reuters, UK
- Barbara Poblete, University of Chile, Chile
- Muhammad Atif Qureshi, National University of Ireland, Ireland
- Paolo Rosso, Universidad Politecnica de Valencia, Spain
- Alan Said, Recorded Future, Sweden
- Damiano Spina, RMIT, Australia
- Jeroen Vuurens, TU Delft, The Netherlands
- Colin Wilkie, University of Glasgow, UK
- Arjumand Younus, National University of Ireland, Ireland
- Arkaitz Zubiaga, University of Warwick, UK

References

1. De Francisci, G., Morales, A.G., Lucchese, C.: From chatter to headlines: harnessing the real-time web for personalized news recommendation. In: Proceedings of WSDM (2012)
2. Mathioudakis, M., Koudas, N.: Twittermonitor: trend detection over the Twitter stream. In: Proceedings of SIGMOD (2010)
3. Aslam, J., Ekstrand-Abueg, M., Pavlu, V., Diaz, F., Sakai, T.: TrREC temporal summarization. In: Proceedings of TREC (2013)