A Framework for Social Semantic Journalism

Bahareh Rahmanzadeh Heravi[1] and Jarred McGinnis[2]

[1]Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway, Ireland
Bahareh.Heravi@deri.org
[2]Logomachy Ltd, London, UK
jarred@logomachy.org

Abstract. Increasingly, news breaks on social media, where ordinary citizens post images and videos and their own commentary in the form of text. This user-generated content (UGC) is newsworthy information and invaluable for newsrooms. In order to incorporate this data into a news story, the journalist needs to process, compile and verify information on the social web within a very short timespan. This is done mostly manually and is a time-consuming and labour-intensive process for media organisations. This paper proposes Social Semantic Journalism framework as an assistant to journalists for breaking news production, and as solution to the above problem.

Keywords: Social Semantic Journalism, Citizen Journalism, User Generated Content, Social News, Semantic News, Social Web, Semantic Web, Linked Data.

1 Introduction

The consumers of news and information are no longer passive and isolated consumers. Smart phones, digital cameras, mobile internet and social media platforms have made us all broadcasters of information. We consume information from traditional news sources, but also through social media platforms, with 1/3 of adults under 30 getting their news from social media [1]. We form communities to inform one another, we comment, we coordinate, and we disseminate. This ubiquity of new technologies has made it more likely than ever that an individual or a community, not a professional journalist, will be the initial source of information for a breaking news event. This community-sourced data, or "citizen/social journalism", is a valuable source of information for news media organisations across the world.

Journalists are already monitoring social media for scoops, details, and images, but the process is laborious, and provides inconsistent results. In the deadline-driven world of journalism, the need to process huge volumes of community-sourced data in order to extract potential news stories is a universal problem. This data, known as user-generated content (UGC) is mostly unstructured, unfiltered and unverified, and often lacks contextual information. Traditional approaches to newsgathering are quickly overwhelmed by the volume and velocity of information being produced.

Extracting stories from UGC goes beyond the simple transcoding of individual streams; it is also important for news organisations to have richly annotated, analysed and interconnected content.

Social Semantic Journalism addresses a universal problem experienced by media organisations; the combination of vast amount of UGC across social media platforms and the limited amount of time the journalist has to spare to extract potential news stories from these mostly unstructured, unfiltered and unverified data. In this situation, there is evidently a need for solutions that can help source, filter and verify social media content for media organisations who are now competing with the continuous flow of free content available 24/7 on the web, while budgets are tight and deadlines are tighter. Social Semantic Journalism also aims to address the chief obstacle facing news organisations, the vetting process, since the current manual process of checking through user-generated content is considered to be overwhelming and inadequate [2].

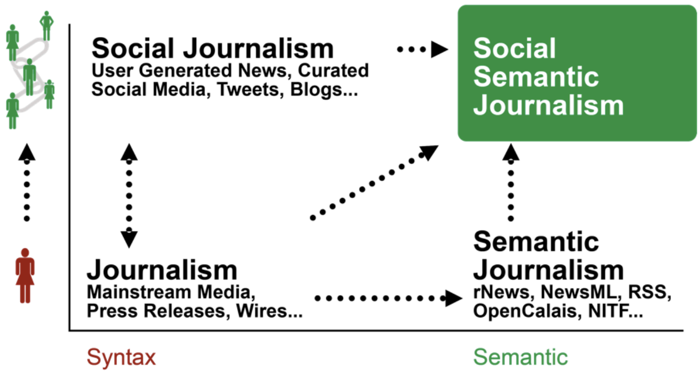
The remainder of this paper is as follows. Section 2 introduces Social Semantic Journalism. Section 3 briefly describes the technologies to best suited to constitute a framework for Social Semantic Journalism. Section 4 concludes the paper with summary remarks about the potential impact of Social Semantic Journalism.

2 Social Semantic Journalism

The user-generated content shared on the social media now forms a significant source of first hand information and content for breaking news coverage. Every minute 347 new blog posts are created, 74 hours of new video is uploaded to YouTube, 100,000 tweets are sent and Facebook users share 684,478 pieces of content [3]. Amongst these data is valuable information that the professional journalist can use to create breaking news stories. This huge amount of user-generated content, however, cannot be processed manually and there is no existing search engine or online tool that can source, aggregate, filter and verify these fast paste and voluminous streams of data for news reportage.

Social Semantic Journalism proposes a Semantic-based solution that can formalise and link unstructured UGC to other semantically-enriched data sets in what is termed the "Linked Data Cloud" for integration, verification and fact-checking purposes, e.g. government datasets or DBpedia/Wikipedia. By working with the media industry, Semantic Web researchers can significantly add to the emerging field of computational or data journalism by "developing techniques, methods, and user interfaces" that can "help discover, verify, and even publish new public-interest stories at lower cost" [5].

Semantic Web technologies are a means for providing a machine readable data structure and also facilitate information integration from various sources which are built using the same underlying technologies. The Semantic Web effort is considered to be in an ideal position to make social web platforms interoperate by providing standards to support data interchange and interoperation [10]. The application of the Semantic Web to the Social Web, termed the "Social Semantic Web", has the potential to create a network of interlinked and semantically enriched user generated knowledge base, bringing together applications and social features of the Social Web with knowledge representation languages and formats from the Semantic Web [10].



Social Journalism
User Generated News, Curated Social Media, Tweets, Blogs...
Social Semantic Journalism
Journalism
Mainstream Media, Press Releases, Wires...
Semantic Journalism
rNews, NewsML, RSS, OpenCalais, NITF...
Syntax
Semantic

Fig. 1. Social Semantic Journalism, adopted from [11]

Figure 1 depicts Social Semantic Journalism as the convergence of technological and cultural trends [11]. Ontologies are at the heart of Semantic Web technologies and provide a formal and semantically enriched description of concepts and their relationships within a domain with the aim of a shared understanding. There are a number of well-defined ontologies in the social web realm such as SIOC (Semantically Interlinked Online Communities) [5] and FOAF (Friend Of a Friend) [7]. On the Semantic Journalism side ontologies such as rNews [8] and SNaP [9] provide semantic markup for annotating news stories with metadata for web documents (rNews) and at the enterprise level (SNaP).

Making use of the these semantic foundations and Linked Data principles, it is possible to develop a framework for Social Semantic Journalism by employing a number of technologies and processes, which is described in the next section.

3 A Framework for Social Semantic Journalism

There are a number of technologies that will be required to produce a Social Semantic Journalism Framework. These technologies would inevitably work together, becoming the inputs and outputs for each other, creating a process up to meet the challenge that social media presents to journalists and editors as they try to what is news worthy in UGC. Figure 2 illustrates the technologies and process to realise Social Semantic Journalism.

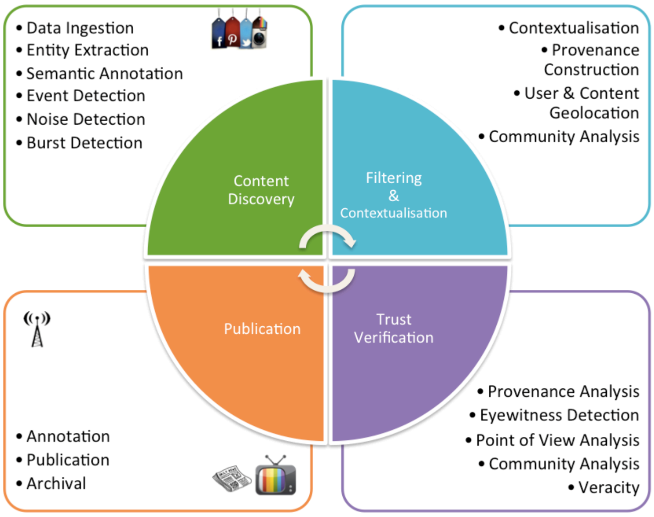


Fig. 2. Social Semantic Journalism Framework

Content Discovery is the ingestion the raw content from social media and enriching it with semantic metadata, which can be made use of by the other phases.

Data Ingestion is gathering a representative sample of data from microblog updates and the users posting them.

Entity Extraction and *Semantic Annotation* involves the identification of semantic entities such as ‘place’, ‘organisation’ and ‘person’ and linking them with relevant semantic metadata from Linked Open Data.

Event Detection is the identification of events as they happen.

Noise Detection is the detecting and filtering non-relevant content from streams where a topic has already been identified.

Burst Detection is the discovery of bursts or sudden increases in frequency of topic and/or location specific microblog.

Filtering and Contextualisation uses the derived metadata from Content Discovery phase, and further refines the metadata, associating related content, putting news stories within a wider context of the news agenda and world events and starting to develop a provenance trace.

Contextualisation discovers background and contextual information for a specific news story, leveraging the metadata created during the content discovery stage.

Provenance Construction is to produce a provenance trace and graph to be utilised for the trust verification stage.

User and Content Geolocation is to approximate the relevant location of an event/tweet by exploiting a combination of explicit GPS coordinates, disambiguation through semantic annotation and making use of social graph data.

Community Analysis relies on the event, burst and noise detection to isolate users generating timely and relevant UGC.

Trust Verification utilises the provenance data and the extracted concepts from Content Discovery and Filtering and Contextualisation.

Provenance Analysis provides the analysis, abstraction and summarisation of provenance information, which would help journalists in identifying eyewitnesses and assessing reputation of the source.

Point-of-View Analysis provides indicators for the perspective or point-of-view of a piece of content to inform the journalist as to the likely perspective the content takes.

Veracity determines the veracity of the content of a post.

Publication is concerned with annotation, publication and archival of produced news stories. This phase feeds back to filtering and contextualisation phase for future historical contextualisation purposes.

4 Conclusions

This paper introduced Social Semantic Journalism framework, as a solution that can help journalists in the process of breaking news production, when the initial source is social media.

The potential impact of a framework for Social Semantic Journalism includes a dynamic and flexible alternative to newswire subscriptions, providing high-quality and timely news and open up the market to new media aggregators, curators and commentators, creating new business opportunities and opportunities for media exploitation and reuse. The novel ability to effectively exploit, large-scale social media streams by journalists will give a voice to citizens, enabling journalists to do richer and more relevant story development faster.

There is increasing evidence of a tipping point for technologies such as semantic web, linked data and natural language processing. Non-technology companies in the news industry sector such as BBC, New York Times, Novosti and the Press Association have begun to make considerable capital investments in the technologies employed with this project (e.g. linked data, language analytics, etc.). The benefits of using Linked Data for knowledge sharing, integration and reuse has been validated in a variety of application domains and contexts from the digital enterprise to healthcare and green IT. A social semantic journalism framework is an opportunity to demonstrate the viability of these approaches for integrating social media information from a community of users into the mainstream news media workflow.

An immediate consequence of the framework would be the development of a set of API-driven semantic services and tools to process social media content and data is the stimulation of demand for high-performance, bandwidth-hungry media applications and services. The torrent of disparate, contradictory and unstructured social media content is made accessible, relevant and useful.

Acknowledgement

The work presented in this paper has been partially funded by Science Foundation Ireland (SFI) under Grant Numbers 2/TIDA/I2389 and SFI/08/CE/I1380 (Lion-2).

Bibliography

1. Sonderman, Jeff. One-third of adults under 30 get news on social networks now. http://www.poynter.org/latest-news/mediawire/189776/one-third-of-adults-under-30-get-news-on-social-networks-now/
2. Rosen, J. 2008. Definition of Citizen Journalism [Online], Available from: http://www.youtube.com/watch?v=QcYSmRZuep4.
3. DOMO, How Much Data is Created Every Minute? Available from:http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/
4. Hussain, M. M, and Howard, P. N. 2010. Opening Closed Regimes: Civil Society, Information Infrastructure, and Political Islam. In Annual meeting of the American Political Science Association.
5. Cohen, S., Hamilton, J.T. & Turner, F., 2011. Computational journalism. Communications of the ACM, 54(10), p.66.
6. Berrueta, D., Brickley, D., Decker, S., Fernández, S., Görn, C., Harth, A., Heath, T., Idehen, K., Kjernsmo, K., Miles, A., Passant, A., Polleres, A., Polo, L. & Sintek, M. (2007). *SIOC Core Ontology Specification* (W3C Member Submission). W3C.
7. Tramp, S., Frischmuth, P., Ermilov, T., Shekarpour, S. & Auer, Sö. (2012). An Architecture of a Distributed Semantic Social Network. *Semantic Web Journal*, Special Issue on The Personal and Social Semantic Web.
8. http://dev.iptc.org/rNews.
9. McGinnis, J., Wilton, P., Harman P., O'Donovan, J. (2012) http://data.press.net/ontology/.
10. Breslin, J.G., Passant, A, Decker, S. The Social Semantic Web: Springer, ISBN 9783642011719, 3 October 2009.
11. Heravi, B. R., Boran, M., & Breslin, J. (2012, May). Towards Social Semantic Journalism. *Workshop on the Potential of Social Media Tools and Data for Journalism in News and Media Industry at the Sixth International AAAI Conference on Weblogs and Social Media.*
12. Avram, A. (2012). Gartner's Software Hype Cycles for 2012. http://www.infoq.com/news/2012/08/Gartner-Hype-Cycle-2012.