AAAI Technical Report WS-12-01
The Potential of Social Media Tools and Data for Journalists in the News Media Industry

Towards Social Semantic Journalism

Bahareh Rahmanzadeh Heravi[1], Marie Boran[1], John G. Breslin[1, 2]
[1] Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, Ireland
[2] School of Engineering and Informatics, National University of Ireland, Galway, Ireland
{Bahareh.Heravi, M.Boran, John.Breslin}@deri.org

Abstract

User-generated content has become a valuable journalistic tool for news coverage and production. This convergence of new and old media, however, poses several challenges to established news organisations. Social media sites produce a wealth of data in the form of text, images and video that must be processed, compiled and verified within a very short timespan before being incorporated into a news story. This unstructured data that lies scattered across the web can be formalised and organised into a 'web of data' by Semantic Web technologies. Specifically, Semantic Web technologies have the potential to formalise and integrate artifacts produced and shared across the Social Web. Social Semantic Journalism proposes the utilisation of Semantic Web technologies, and specifically Social Semantic Web ontologies such as FOAF and SIOC, in the process of news production. This potentially provides a journalistic tool to assist in finding, aggregating and verifying user-generated content for news production.

Introduction

In mainstream journalism the traditional news gathering process is defined by the methods employed by the profession itself, one that uses "certain techniques of news-gathering and construction" while seeking information "in official places" (Hindman 1998, p. 177) such as press releases, the news wire and first person interviews. Social media platforms and user-generated content have changed this process; the audience is now a readily available source of news.

Rosen (2008) refers to citizen journalism as "the people, formerly known as the audience", who use social media platforms such as Twitter and YouTube to inform one another. News organisations have begun to adapt to this information flow by incorporating social media into the news gathering process. Examples are CNN's iReport and the BBC, which has established a dedicated section for the material sent by the public, called the UGC (User Generated Content) hub, in the heart of its newsroom.

While social media produces a vast amount of user-generated news, the content comes from multiple authors and multiple platforms and a considerable amount of time is required to find, compile and verify this content. Semantic Web technologies provide a machine readable data structure and facilitate information integration from various sources and are considered an ideal solution for making Social Web platforms interoperate. This paper seeks to investigate the ways in which Semantic Web technologies may be used in conjunction with the Social Web for finding, compiling, aggregating and validating newsworthy material posted on the Social Web by content-generating users/citizen journalists, taking inspiration from the areas of opportunity in computational journalism proposed by (Cohen et al. 2011).

This paper is organised as follows. The next section provides an overview of the ways in which social media has changed news production as well as consumption in the recent years. It also studies various sources which may be considered in the process of news production and the challenges associated with this process. The third section provides a brief overview of Semantic Web and ontologies and provides a brief review on the relevant work in the news and journalism domain. The fourth section introduces the Social Semantic Web as an approach for creating a network of interlinked and semantically enriched user-generated content. It further introduces Social Semantic Journalism, which proposes to use Semantic Web technologies to address the challenges associated with news gathering and verification in the Social Web. It then provides examples of the ways in which Social Semantic Journalism is seen to be implemented. Finally the last section summarises the paper and provides a set of future research directions.

Journalism and the Social Web

Social media platforms have evolved from being seen as an alternative channel by which to market content to the audience to a legitimate source of news provided by citizen journalists (Lowery 2009). The Arab Spring, as an event of global political significance and media coverage, was defined in part by its coverage on Twitter and YouTube. This produced huge amounts of user-generated content that

became a valuable source of news for the mainstream media (Hussain and Howard 2011).

Accordingly major news broadcasters have recently began investing in the utilisation of social media. According to Journalism.co.uk, social media ranks top of news media investment in the next five years (McAthy 2012). Major news and media broadcasters nowadays use social media and user-generated content as an important source for gathering news, particularly from places where there are no journalists, e.g. geographically remote/difficult regions. In such circumstances, the news comes from variety of social media sources, such as Twitter, Facebook, Google+, YouTube, various blogs or the content submitted directly to the broadcasters web services, such as CNN iReport. Figure 1 presents the sources which may be compiled in the event of a breaking news in order to come up with a trustworthy piece of news.

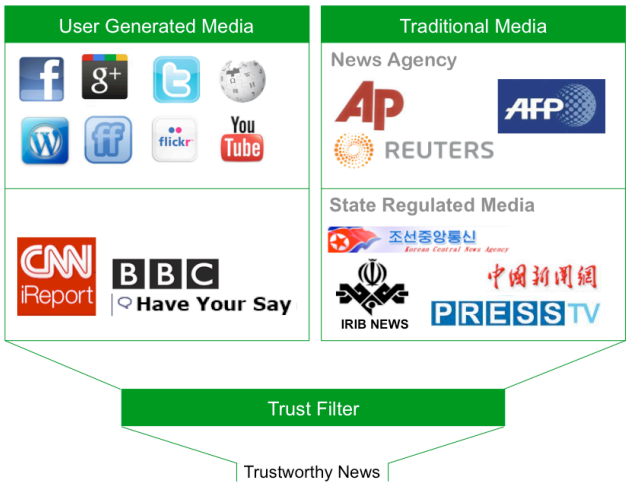

Figure 1. News sources and trust as a part of the news production process

Navigation of data may be considered as a major challenge in the Social Web, where platforms are not connected and communities are often disparate. The social media sources are not well connected/integrated and the UGC producers would normally have to spend a considerable amount of time searching, compiling and verifying the material received on those sources. In the other words there is a huge amount of user-generated content available on the (social) web, which is not connected together and acts as a set of separate data silos that cannot interoperate and do not understand each other (Breslin, Passant and Decker 2009). Furthermore, this information lacks exchangeable semantics and thus is not as usable as it can be for creating trustworthy news stories.

Another challenge for a news organisation in blending traditional and participatory journalism is the vetting process. Sambrook (2005, p. 14) describes the manual process of checking through user-generated content as overwhelming and "inadequate as we go forward". CNN's solution to verifying external content is to have a registration process. While this is adequate for the iReport site, it does not address the issue of content on social media platforms like Twitter, Facebook, Google+ or YouTube.

Journalism and the Semantic Web

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. Semantic Web technologies are believed to provide considerable improvements in the way news materials are gathered from a variety of social media sources.

Traditional media organisations are already embracing Semantic Web technologies, but not yet in combination with the user-generated content. The BBC is an example of a broadcasting corporation which is utilising Semantic Web technologies in their web sites, mainly for their routine programs and music, and not yet for news. Another example is Thomson Reuters, which has established a subsidiary called ClearForest that endeavors to connect related information across different platforms without the need for an editor to oversee these actions by using Semantic Web technologies (Sanborn 2008). In 2008 ClearForest released Calais, a toolkit that unlocks Semantic Web functionality across various online platforms including blogs and content management systems.

A number of academic projects also exist which have considered the use of semantic web technologies in the news production life cycle. Troncy (2008) suggests that ontology should be used in the news workflow process in order to reduce the interoperability problems caused by using different metadata formats within the news production chain and also for improving and facilitating the search and brows of news content for end users. They design an OWL ontology for IPTC (International Press Telecommunications Council) News Architecture Framework (NAR[1]). The NEWS Project (Zapf, Fernandez-Garcia and Sanchez-Fernandez 2005) is another relevant initiative aiming at bringing the Semantic Web technologies into the news industry with the aim of helping the news agencies to improve their news production and distribution processes. An ontology for the news domain is implemented as part of the NEWS project, taking into account the existing standards in the news industry, mainly the IPTC standards (Fernández, Fuentes, Sánchez, Fisteus 2010). News@hand (Cantador, Bellogín and Castells 2008) is another ontology based news system, aiming at porducing enhanced news recommendations by describing

[1] http://www.iptc.org/NAR/

and relating news contents, retrived from RSS feeds of various online news services, and user preferences.

The Semantic Web tecnologies have also been used in the process of decision making about the relevant news stories and items. An example is the work of Borsje, Levering and Frasincar (2008), who propose an ontology based decision support framework, HERMS, for helping decision makers to extract news items which are related to a specific topic on interest by analysing, categorising and visualising the data coming from the RSS feeds published from vasious sources, e.g. news agencies.

The above work however, mainly focus on the post story production, i.e. storage and retrival of produced news items, and do not take advantage of the Semantic Web for gathering, processing and verifying and producing news stories. This paper focuses on the task of assisting journalists with content creation by providing them contextual information for aggregating and verifying user generate news and making judgments related to a social media item or a media item from an unknown source.

Social Semantic Journalism

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. 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. Various vocabularies, or ontology languages, can be used to interlink and aggregate Social Web content in the context of news and journalism. Figure 2 depicts Social Semantic Web Journalism and its underlying concepts as proposed in this paper.

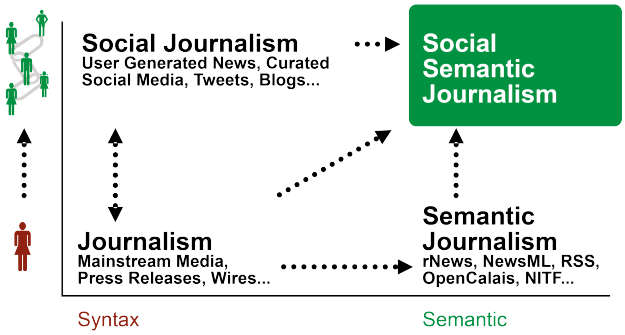

Figure 2. Social Semantic Journalism

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 realm of the Social Semantic Web, which could be used to uniformly represent the different artifacts produced and shared in the Social Web: communities, people, documents, tags, etc. The most popular of these ontologies are SIOC (Semantically-Interlinked Online Communities) and FOAF (Friend Of a Friend).

SIOC aims at interconnecting Social Web platforms, enabling the integration of online communities information by providing an ontology for representing rich data from the Social Web in RDF. By becoming a standard way for expressing user-generated content from social web sites, SIOC enables new kinds of usage scenarios for online community site/user-generated data, and allows innovative semantic applications to be built on top of the existing Social Web. One such scenario is Social Semantic Journalism, which would enable more advanced methods for the extraction, verification and compilation of user-generated content across various social media platforms. rNews is another relevant journalistic initiative, providing semantic markup for annotating news specific metadata in web documents. rNews is an approved standard, developed by the IPTC, a consortium of the world's major news agencies, news publishers and news industry vendors.

Sourcing relevant social media content would form an important ingredient of the Social Semantic Journalism and one way to do so is to determine the importance of the content by context. Sindice (Decker, Delbru and Polleres 2008) is a semantic index of the Web, which allows finding pointers to relevant pages or URIs where particular keywords are mentioned, where certain property values are used, or where certain facts or semantic triples appear. The application of Sindice, i.e. finding pointers to things, and using that in Social Semantic Journalism, could potentially be very powerful.

An example of the above application is that Sindice can be used to browse a combination of distributed SIOC documents (related to a person or a topic) via the Sindice index. Therefore, if you are browsing a social media item made by a particular person, you could see a pop-up window with a list of content (posts, comments, topics) that that person has created not just on the site you are viewing but across a range of SIOC-enabled websites (Twitter, Facebook, blogs, forums, etc.) as indexed in Sindice. You could then navigate to the content a person has created across a range of sites from just one place that they post to, to obtain some more context about that person and what they tend to write about (this is enabled through a FOAF person that holds multiple online accounts). Alternatively one could find content on similar topics to the current item from a distributed set of sites.

Sindice also has an API that can provide results in a reusable (semantic) format that can be leveraged by other

applications, and it powers Sig ma, a semantic mash-up application that allows users to type in a keyword and view information about the entity that best matches that keyword as mashed up from a variety of sources. In addition, semantic data about a user account (e.g. bio interest, Klout influence levels and lists that a user is mentioned in), could be used to help the journalists with background checks.

The focus of this project is to explore the ways in which Semantic Web technologies and ontologies, such as SIOC and FOAF, in conjunction with news representation standards, such as rNews, could assist in the process of interlinking online user communities and the user-generated content for news gathering and verification. Likewise, it would consider extending/customising SIOC for its utilisation in journalism and the news industry. This work would take into consideration the existing work on the use of the Social Web in the media industry (such as Storyful and NewsWhip) and also the standardisation work such as rNews. This work does not intend to re-invent the wheel, but to use best practice in the realm of the Semantic and Social Web in conjunction with journalism, news and the media industry for achieving a more coherent way for newsgathering and verification.

Future Research Directions

This paper presents the importance of user-generated content in modern journalism and the ways in which the Social Web has affected this industry, its benefits and the challenges associated with it. This research proposes a novel approach for user-generated news production that makes use of Semantic Web technologies and ontologies for formalising, integrating and aggregating newsworthy user-generated content. In order to achieve the goal of the proposed research the following research directions are considered important to be addressed in the future:

- Conducting a comprehensive literature review on the current utilisation of the Social Web within the news industry, both by mainstream broadcasters and independent service providers.
- Conducting a comprehensive literature review on the ways in which Semantic Web technologies have been implemented to date within the news industry.
- Evaluating existing relevant ontologies and data standards within a journalistic context.
- Determining what existing Semantic Web tools can be integrated into this project.
- Defining a set of contextual factors with which to assess trust levels for both social media sources (the individual) and content.
- Proposing a standard for integration and display of subsequently semantically labeled social media content to both journalists and the news audience.

References

Breslin, J.G., Passant, A, Decker, S. *The Social Semantic Web*: Springer, ISBN 9783642011719, 3 October 2009.

Cantador, I., Bellogín, A. & Castells, P., 2008. News@hand: A Semantic Web Approach to Recommending News. In W. Nejdl et al., eds. *Adaptive Hypermedia and Adaptive WebBased Systems*. Springer Berlin / Heidelberg: 279-283.

Cohen, S., Hamilton, J.T., Turner, F. 2011. Computational Journalism, *Communications of the ACM* 54(1): 66-71.

Decker, S., Delbru, R., Polleres, A. and Tummarello, G. 2008. Context Dependent Reasoning for Semantic Documents in Sindice, *Science*: 23.

Fernández, N., Fuentes, D., Sánchez, L. Fisteus, J. A. 2010, The NEWS ontology: Design and applications, *Expert Systems with Applications,* 37(12): 8694-8704.

Hindman, E. B. 1998. Spectacles of the Poor: Conventions of Alternative News, *Journalism and Mass Communication Quarterly* 75(1): 177-193

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*.

Jensen, E. 2010. Between credulity and scepticism: envisaging the fourth estate in 21st-century science journalism. *Media, Culture & Society* 32(4): 615-630.

Lowery, C. 2009. An Explosion Prompts Rethinking of Twitter and Facebook [Online], Niemen Reports. Available from: http://www nieman harvard.edu/reportsitem.aspx?id=101894 [Accessed 29 February 2012].

McAthy, R. 2012, Report: Social media top for future news outlet investment [Online]. Available from: http://www.journalism.co.uk/news/study-finds-majority-of-news-outlets-rank-social-media-top-for-investment-in-next-5-years/s2/a548068/ [Accessed 07 March 2012].

Oriella PR Network 2011. *The State of Journalism in 2011: Oriella PR Network Digital Journalism Study*.

Rosen, J. 2008. Definition of Citizen Journalism [Online], Available from: http://www.youtube.com/watch?v=QcYSmRZuep4 [Accessed 29 February 2012].

Sanborn, J. 2008. Welcome to Web 3.0, *American Journalism Review*. Available from: http://ajr.org/Article.asp?id=4604 [Accessed 29 February 2012].

Troncy, R., 2008. Bringing the IPTC News Architecture into the Semantic Web A. P. Sheth et al., eds. *Lecture Notes In Computer Science* Vol 5318: 483-498.

Zapf, L., Fernandez-Garcia, N.and Sanchez-Fernandez, L., 2005. The news project - semantic web technologies for the news domain, *2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, London.*