Research Brief:

Introducing Social Semantic Journalism

Bahareh Rahmanzadeh Heravi

Insight Centre for Data Analytics National University of Ireland, Galway, Ireland

Bahareh.Heravi@insight-centre.org

Jarred McGinnis

Logomachy Ltd, London, UK

jarred@logomachy.org

ABSTRACT

In the event of breaking news, a wealth of crowd-sourced data, in the form of text, video and image, becomes available on the Social Web. In order to incorporate this data into a news story, the journalist must process, compile and verify content within a very short timespan. Currently this is done manually and is a time-consuming and labour-intensive process for media organisations. This paper proposes *Social Semantic Journalism* as a solution to help those journalists and editors. Semantic metadata, natural language processing (NLP) and other technologies will provide the framework for Social Semantic Journalism to help journalists navigate the overwhelming amount of UGC for detecting known and unknown news events, verifying information and its sources, identifying eyewitnesses and contextualising the event and news coverage journalists will be able to bring their professional expertise to this increasingly overwhelming information environment. This paper describes a framework of technologies that can be employed by journalists and editors to realise Social Semantic Journalism.

KEYWORDS:

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

The Journal of Media Innovations 2.1 (2015), 131–40.

http://www.journals.uio.no/index.php/TJMI

INTRODUCTION

Smart phones, digital cameras, the mobile internet and social media platforms have made us all broadcasters of information. The 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/denizen journalism", is a valuable source of information for news media organisations across the world.

'Citizen Journalism' is an oft-used term in media studies and as it continues to evolve, it is difficult to put settled boundaries on the term (Lasica, 2003). Various definitions for 'citizen journalism' could be found in the literature (e.g. Robinson, 2006; Nip, 2006; Rosen, 2008, Thurman, 2008, Goode, 2009; Robinson and DeShano, 2011; Örnebring, 2013). For the purpose of this study we consider a citizen journalist as an ordinary citizen who provides eyewitness commentary on a current event, either intentionally or accidentally. Under the current technological climate most of the content produced by citizen journalists may be shared on social media websites, but we do not believe the definition should be limited to content produced and/or shared on such media websites as it may be circulated via other communication platforms or apps. Furthermore, our definition does not limit citizen journalists to only those who are not paid (unlike the view of Robinson (2006)); We believe citizen journalists are not employees of news/media organisations, but may be paid through sporadic crowdsourcing calls.

Despite all the arguments, there are examples where 'citizen journalists' have been able to set the news agenda and break stories; moreover, there are many well-known examples of media content that would never have appeared on mainstream media if it weren't for citizen journalists. In the June of 2009, arguably the most famous of early user-generated videos to date was the footage of Neda Aghda Soltan's death during the protests following the 2009 Iranian elections. A video showing her death from a single gunshot wound to the chest was uploaded on the internet. Later, two other videos of her were published on the internet, which covered her horrific death from different angles. Neda Aghasoltan's death was then referred to as "probably the most widely witnessed death in human history"(Mahr, 2009). The Mumbai 2008 bomb blasts, the 2011 crash of US Airways Flight 1549, the Arab Spring movements, and the Boston Marathon bombing are other widely known examples that spurred user-generated content, i.e., as generated by citizen journalists.

The literature suggests that the citizen journalist has become an integral part of the media and her/his role and aims have, arguably, become similar to those of any journalist. This has also changed the way traditional journalists "scoop" the news: A survey of journalists in fifteen countries showed that half the respondents used Twitter to source angles for a news story (Oriella PR, 2013). This introduces new challenges to the industry, and calls for new, innovative business models to address those challenges.

Journalists are already monitoring social media for scoops, details, and images. This, however, is largely a manual process, which 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 – namely, the combination of a vast amount of UGC across social media platforms and the limited amount of time that 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 that 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 (Rosen, 2008).

A Social Semantic Journalism (SSJ) framework will:

- Increase the potential for the social media content to be discovered, used, shared and integrated in automatic ways as the pillar of Social Semantic Journalism;
- Maximise the opportunities to capture relevant user generated content for discovery, and to automatically enrich it with related knowledge including geographical context, semantic relations and social interactions, facilitating filtering and aggregation in an interoperable way;
- Provide richer multi-dimensional ways of assessing the quality of the content by combining knowledge and metrics from the social space and the semantic relatedness of content and users in a timely manner;
- Facilitate sharing, integration and reuse of news in digital journalism, by applying linked data principles to publication and Archival for better interoperability.

We define Social Semantic Journalism in Section 2. Section 3 describes the technologies best suited to constitute a framework for Social Semantic Journalism and deliver the potential efficiencies in the news production process, enabling enhanced news searches and providing a finer understanding of the information an organisation consumes and produces. The paper concludes with summary remarks about the potential impact of Social Semantic Journalism.

SOCIAL SEMANTIC JOURNALISM

User-generated content shared on social media plays a significant role in the process of capturing news events, classifying and verifying stories and also keeping the audience in the loop with timely and accurate news. Every minute over 350 new blog posts are created (James, 2012), 100 hours of new video is uploaded to YouTube (YouTube, 2014), over 540,000 tweets are sent (Statisticbrain, 2014) and Facebook users share 684,478 pieces of content (James, 2012). Amongst these data is valuable information that the professional journalist can use to create breaking news stories – but this cannot all be processed manually and there is no existing search engine or online tool that can source, aggregate, filter and verify this content 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" (Cohen, 2011).

Semantic Web technologies are a means for providing a machine-readable data structure and also facilitating 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 (Breslin, 2009). 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 (Breslin, 2009).

Figure 1 depicts Social Semantic Journalism as the convergence of technological and cultural trends (from Heravi, 2012). Ontologies are at the heart of Semantic Web technologies and provide a formal and semantically enriched description

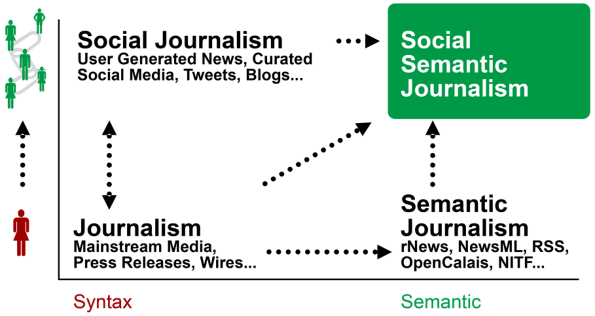
Social Journalism
User Generated News, Curated Social Media, Tweets, Blogs...
Social Semantic Journalism
Journalism
Mainstream Media, Press Releases, Wires...
Semantic Journalism
rNews, NewsML, RSS, OpenCalais, NITF...
Syntax
Semantic

Figure 1. Social Semantic Journalism

of concepts and their relationships within a domain with the aim of shared understanding. There are a number of well-defined ontologies such as SIOC (Semantically Interlinked Online Communities) (Berrueta, et al, 2009) and FOAF (Friend Of a Friend) (Tramp et al, 2012). For example, 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. FOAF is an ontology for describing people and the links between them. For Semantic Journalisms, there are ontologies such as schema.org[1], which is used for representing news articles on web documents; SNaP (McGinnis, 2012), which is used for representing news concepts within news for an enterprise content management system; and the BBC's Storylines[2], which is used for organising news content to broad thematic collections and narratives. 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, as described in the next section.

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 newsworthy in UGC. Figure 2 illustrates the technologies and process to realise Social Semantic Journalism.

Content Discovery

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. Data ingestion includes the contents of the post, any re-post information, mentions of other users, the URI of the post, any links from the post, the timestamp of the post and any explicit location information (geo tags) attached to the post. Additional location information will be derived from location-based network analysis, entity extraction and semantic annotation.

[1] http://schema.org/NewsArticle

[2] http://www.bbc.co.uk/ontologies/storyline/

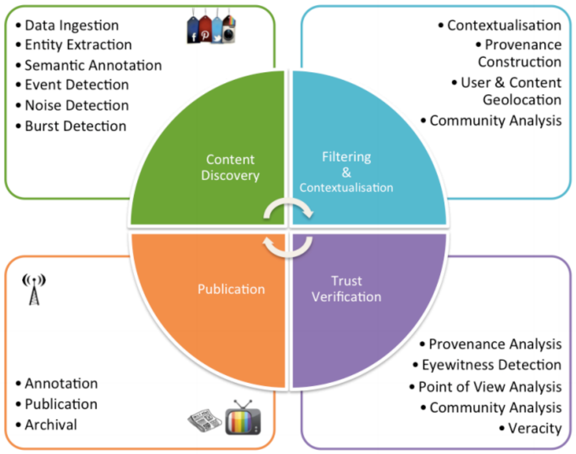

Figure 2. Social Semantic Journalism Framework

Entity Extraction and *Semantic Annotation* involves employing Natural Language Processing (NLP) techniques for building a semantic annotation tool to correctly identify mentions of entities in social media streams and to further link them with relevant semantic metadata from Linked Open Data for later indexing and search. Ontology-based Information Extraction techniques will assist in exploiting the hierarchical structure of an ontology to inform the entity extraction task. Three main stages are involved within this task: (a) detection of entities in the text (including specification of entity types); (b) search for potential matches in Linked Open Data; and (c) entity ranking to find the most appropriate entity to link the mention to.

Event Detection is the identification of semantic entities such as (“the queen opening parliament”, “car bombing”, “Manchester United vs. Liverpool”) from streams of disparate data sources such as blogs, microblogs, and comments in breaking news stories from news organisation websites. There are a number of techniques which can be applied; the most straightforward is a stream processor that extracts named entities and hashtags from streaming UGC and monitors the frequency acceleration of these terms over time.

Noise Detection is the detecting and filtering non-relevant (or noisy) content from streams where a topic has already been identified. Principles from information retrieval and the theory of relevance could be applied to produce an approach that ranges from completely automatic with no user involvement to semi-automatic where limited involvement from the end user will be solicited.

Burst Detection is the discovery of bursts or sudden increases in frequency of topic and or location specific microblog events in order to identify when a particular event or news story is trending. It is important to Social Semantic Journalism to know when events start to trend but also about bursts of opinion within an event, which requires more complex analysis.

Filtering and Contextualisation

Filtering and Contextualisation uses the derived metadata from Content Discovery phase, and further refines the metadata by associating related content, putting news stories within a wider context of the news agenda and world events and starting to develop a provenance trace. It will also develop a computational model of newsworthiness to offer better story leads automatically to the journalist users.

Contextualisation discovers background and contextual information for a specific news story, leveraging the metadata created during the content discovery stage. This can be achieved utilising LOD (Linked Open Data) sources, moderated and trusted archive and repositories, word sense, topic disambiguation and similarity-based query processing for context retrieval.

Provenance Construction is achieved by combining information diffusion techniques and abductive reasoning. This produces a provenance trace and graph to be utilised for the trust verification stage.

User and Content Geolocation must be done to address the problem of geolocating users and user-generated content. This is a combination of exploiting explicit GPS coordinates (found in approximately 20% of tweets), disambiguation through semantic annotation and making use of social graph data.

Community Analysis can help with the filtering of events. This task relies on the event, burst and noise detection to first isolate users generating timely and relevant UGC. A graph can be constructed over the users to determine the number of communities, their size, the strength of the connection amongst users and other higher-order properties to provide information about the nature of the communities.

Trust Verification

Trust Verification utilises the provenance data and the extracted concepts from Content Discovery and Filtering and Contextualisation. Trustworthiness, veracity, bias detection within the context of community of users is part of the verification stage

Provenance Analysis provides the analysis, abstraction and summarisation of provenance information. Provenance indicators that would be useful for journalists include identifying eyewitnesses, summarising the overall quality of the provenance trace and whether content comes from a reputable source.

Point-of-View Analysis provides indicators for the perspective or point-of-view of a piece of content, by combining work on opinion mining, bias detection and community detection to inform the journalist as to the likely perspective the content takes and thus aid the journalist in better judging the content.

Veracity can be determined by creating indicators for veracity, using statistical analysis of digital content and quantitative methods as preliminary filters coupled with formal methods in a knowledge-driven fashion to produce more finely grained veracity assessment.

Publication

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.

IMPACT AND CONCLUSIONS

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 as well as opening up the market to new media aggregators, curators and commentators, thereby creating new business opportunities and opportunities for media exploitation and reuse. The novel ability for journalists to effectively exploit, large-scale social media streams without depending on expensive 'fire-house' subscriptions will give a voice to citizens, enabling journalists to do richer and more relevant story development faster.

As social media users see their content broadcasted by mainstream news providers, the power of being an active eyewitness of an event will be obvious. The interaction between professional journalists and active producers of User-Generated Content (UGC) will encourage the adoption of the guiding ethical principles of journalists with whom they interact. This begins a virtuous cycle as the quality and importance of UGC spreads from early adopters within the newsroom to entire news organisations as an effective means of information gathering.

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 the BBC, the New York Times, Novosti, The Financial Times 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.). According to Gartner's Software Hype Cycles for 2012 (Avram, 2012), text and social analytics and Big Data will reach maturity (the so called 'Plateau of Productivity'), within 2 to 5 years with the Internet of things, crowdsourcing, automatic content recognition and complex event processing expected in the midterm. 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.

Social Semantic Journalism facilitates new ways to interact with social media data by aiming at non-expert users, for example freelance journalists. This "democratises" the analytics of data, specifically social media data, and empowers the users to extract value. Journalists will be able to explore, monitor and interface with data such as events on diverse levels of granularity and depth. Filters and facets and intuitive visual cues that could be based on the framework could aid the journalist to navigate and explore complex topics and data streams from communities, biased or otherwise, to narrow the focus of a story and determine the level and veracity of detail to deliver the answers.

The implementation of the framework as a set of APIs (Application Program Interface) would provide the tools to transform the torrent of disparate, contradictory and unstructured social media content into content that is more accessible, relevant and useful.

The technologies described are not aimed at replacing the expertise of journalists. The knowledge, experience and intuition of journalists are too fine to be replaced by current technologies. Instead the goal is to create tools that mediate the inhumane amounts of data and content that are created on a daily basis (i.e., reducing the signal to noise ratio). The decision to 'publish or spike' remains with editorial staff. For example, nuanced usage of language such as irony, satire and sarcasm are notoriously difficult to detect by automated means.

However, at this very moment, the Semantic Web and Linked Data technologies in conjunction with the Social Web are beginning to empower the professional journalist, whose role as a watchdog for democracy is vital to the public interest. Most, if not all, of the aspects of the Social Semantic Journalism Framework described in this paper will play a part in that.

REFERENCES

Avram, A. (2012). *Gartner's Software Hype Cycles for 2012*. Available from: http://www.infoq.com/news/2012/08/Gartner-Hype-Cycle-2012 (Last accessed August 2014).

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.

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

Cohen, S., Hamilton, J.T. & Turner, F., (2011). Computational journalism. Communications of the ACM, *54*(10), 66.

Goode, L. (2009). Social news, citizen journalism and democracy. *New Media & Society*, *11*(8), 1287–1305.

Heravi, B. R., Boran, M., & Breslin, J. (2012). Towards Social Semantic Journalism. Paper presented at Workshop on the Potential of Social Media Tools and Data for Journalism in the News Media Industry at the *Sixth International AAAI Conference on Weblogs and Social Media*. Dublin, Ireland.

Hussain, M. M, and Howard, P. N. (2010). Opening Closed Regimes: Civil Society, Information Infrastructure, and Political Islam. Paper presented at the *Annual meeting of the American Political Science Association*. Washington, D.C.

James, Josh (2012). *DOMO, How Much Data is Created Every Minute?* Available from:http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/ (Last accessed August 2014).

McGinnis, J., Wilton, P., Harman P., O'Donovan, J. (2012). SnaP Ontologies [online], Available from http://data.press.net/ontology/ (Last accessed August 2014)

Nip, J. Y. M. (2006). Exploring the Second Phase of Public Journalism. *Journalism Studies*, *7*(2), 212–236.

Örnebring, H. (2013). Anything you can do, I can do better? Professional journalists on citizen journalism in six European countries. *International Communication Gazette*, *75*(1), 35–53.

Oriella. (2013), *The New Normals for News: Have Global Media Changed Forever?* Oriella PR Network Global Digital Journalism Study 2013. Available from http://www.oriellaprnetwork.com/sites/default/files/research/Brands2Life_ODJS_v4.pdf (Last accessed June 2014)

Robinson, S. (2006). The mission of the j-blog: Recapturing journalistic authority online. *Journalism*, *7*(1), 65–83.

Robinson, S., & DeShano, C. (2011). “Anyone can know: Citizen journalism and the interpretive community of the mainstream press”. *Journalism*, *12*(8), 963–982.

Rosen, J. (2008). *Definition of Citizen Journalism*. Available from: http://www.youtube.com/watch?v=QcYSmRZuep4.

Statisticbrain. (2014). Twitter Statistics. Available from http://www.statisticbrain.com/twitter-statistics/ (Last accessed August 2014).

Sonderman, Jeff (2012). *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/

Thurman, N. (2008). Forums for citizen journalists? Adoption of user generated content initiatives by online news media, *New Media & Society*, *10*(1), 139–157.

Tramp, S., Frischmuth, P., Ermilov, T., Shekarpour, S. & Auer, S. (2012). An Architecture of a Distributed Semantic Social Network. Semantic Web Journal, *1*(0)

YouTube (2014). *YouTube statistics*. Available from http://www.youtube.com/yt/press/statistics.html (Last accessed 25 August 2014).