Where is the News Breaking? Towards a Location-based Event Detection Framework for Journalists

Where Is the News Breaking? Towards a Location-Based Event Detection Framework for Journalists

Bahareh Rahmanzadeh Heravi[1], Donn Morrison[2], Prashant Khare[1], and Stephane Marchand-Maillet[3]

[1] Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, Ireland
Bahareh.Heravi@deri.org, Prashant.Khare@deri.org
[2] Norwegian University of Science and Technology, Trondheim, Norway
donn.morrison@idi.ntnu.no
[3] University of Geneva, Switzerland
Stephane.marchand-maillet@unige.ch

Abstract. The rise of user-generated content (UCG) as a source of information in the journalistic lifecycle is driving the need for automated methods to detect, filter, contextualise and verify citizen reports of breaking news events. In this position paper we outline the technological challenges in incorporating UCG into news reporting and describe our proposed framework for exploiting UGC from social media for location-based event detection and filtering to reduce the workload of journalists covering breaking and ongoing news events. News organisations increasingly rely on manually curated UGC. Manual monitoring, filtering, verification and curation of UGC, however, is a time and effort consuming task, and our proposed framework takes a first step in addressing many of the issues surrounding these processes.

Keywords: Event Detection, Location extraction, Citizen Journalism, User Generated Content, Social News, Semantic News, Social Web, Semantic Web, Linked Data, Social Semantic Journalism.

1 Introduction

Social media platforms have recently become a prominent mode of sharing real-time information and in doing so have evolved into more than simply a user-to-user interaction medium, but an important asset to widespread source of newsworthy information being circulated every second. This has turned the former consumers of news and information - the audience - into potential broadcasters of breaking news.

The ubiquity of mobile technology combined with social media 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 organisations.

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 for extracting potential news stories is a universal problem. This data, known as

C. Gurrin et al. (Eds.): MMM 2014, Part II, LNCS 8326, pp. 192–204, 2014.

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.

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 [7], 100 hours of new video is uploaded to YouTube [36], over 540,000 tweets are sent [29] and Facebook users share 684,478 pieces of content [7]. Hidden amongst this data is valuable information that the journalist can use to create breaking news stories. However, the scale of the data precludes manual processing and there exist no effective tools that can source, aggregate, filter and verify this content for news reportage.

Detection of newsworthy events is an area of research which can be readily applied to the early stages of the journalistic lifecycle. In the past, event detection from unstructured text has been used for applications from first story detection (FSD) [16], where novel news stories are detected from news organisations, to the more general sense of discovering anomalous patterns within large streams of data [20, 11]. In the journalistic context, event detection aims to decrease the time between the occurrence of a news event and the point at which a journalist is made aware of the event. Location-based event detection would then act as a geographic filter, further decreasing this time span by considering only UGC events occurring in areas where breaking news may be expected.

Figure 1 shows how, in the wider context, the framework described in this paper fits into the emerging topic of Social Semantic Journalism [18, 10], which 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 extract potential news stories from these mostly unstructured, unfiltered and unverified data. In this situation, there is evidently a need for solutions

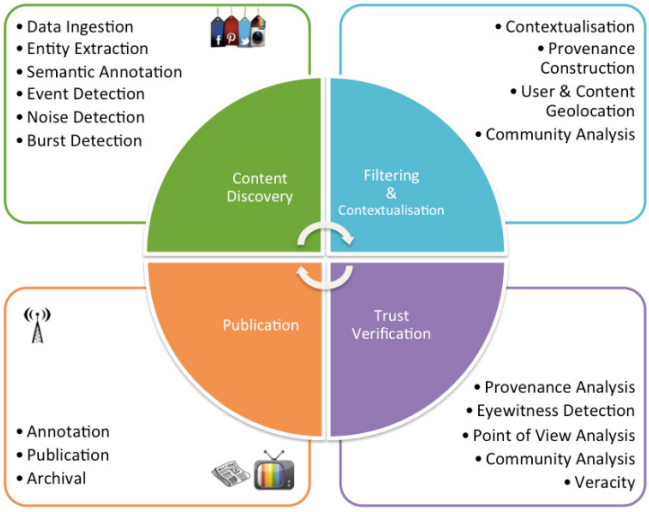

Fig. 1. Social Semantic Journalism Framework, adopted from [10]

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 on the web. Social Semantic Journalism also aims to address the chief obstacle facing news organisations: the vetting process, since the current manual process of verifying user-generated content is considered to be overwhelming and inadequate [21].

In this position paper we focus on the following objectives which we envisage a location-based event detection framework will have in relation to journalism involving UGC:

- Improve access to location-specific events from UGC;
- Decrease the time delay between the event and the reporting of the event by news organisations;
- Aid journalists in assessing the veracity of events reported in UGC; and
- Aid in efforts to trace back to first person reports.

The remainder of this paper is organised as follows. Section 2 describes the related work and existing approaches that could be adapted as components of the framework, both for unstructured UGC as well as associated metadata. Section 3 presents our proposed framework and examines the role of event detection in the journalistic lifecycle. Finally, Section 4 offers concluding remarks on the feasibility and impact of a location-based event detection framework in journalism.

2 Related Work

In the event detection literature, an event is defined as a real-world occurrence with an associated time period and a specific location. Considering the existence of a time-ordered stream of published messages relating to the real-world occurrence, the goal of event detection is to detect the occurrence based on the stream of messages [3]. Classical event detection algorithms can be broadly classified into two categories: message-pivot methods and feature-pivot methods. Message-pivot approaches detect events by clustering messages based on the semantic distance between them. An example is single-pass clustering algorithms [31,1].

A specialised form of message-pivot event detection is first story detection (FSD). FSD in a stream-based setting stores a stream of news stories, each represented as a vector of terms, and compares a new story, i.e., one not yet stored, to all stored stories. Those sufficiently different (by some distance metric or similarity measure such as cosine distance) are flagged “first stories”. The FSD approach can be used to detect events as well as eyewitnesses to help journalists to identify the credibility of the tweets in order to report breaking news and this approach is used in [16]. Feature-pivot approaches involve studying the distributions of words in the messages and discovering events by grouping words together. Examples include Event Detection with Clustering of Wavelet-based Signals (EDCoW) [32] and algorithms for defining communities of keywords. The latter creates a keyword graph of documents or messages and uses community detection methods analogous to those used for social network analysis to discover and describe events [25]. Chen et al. [5] proposed a semi supervised system that crawled the data, specific to organisations and related users, using *fixed keywords* particular to the organisation, its key brands, and prominent

people such as CEO. They developed a classifier which detected the temporally emerging topics from within the *fixed keywords* crawled data and formed the clusters of emerging topics. From emerging topic clusters, a supervised system detects those topics which are fast emerging based on certain features and can be considered as *hot emerging topics*.

Apart from analysing the temporal arrangement of words, there is a lot more information that can be retrieved from the text in terms of context of the messages. Natural language processing (NLP) has been used for event detection from an information extraction (IE) perspective [8]. While the use of NLP encourages the extraction of entities enclosed in text, a supervised learning approach can extract more information about entities and context. TwiCal is presented in [20] as the first open domain event extraction and categorisation tool for Twitter. The system is based on an annotated corpus of events in Twitter which are used as training data for sequence level models. Shallow linguistic analysis is used to create features for the classifier, thereby recognising event triggers as a sequence labeling task, using conditional random fields (CRF) for learning and inference.

Finally, event detection has been explored for other domains such as sports [34, 17, 13, 35, 27, 22] for purposes such as extracting highlights [22], detecting notable events such as point scores [35] and automatically structuring sports video [13]. However, these domains have a considerably smaller scale compared to that available from social media sources.

From a journalistic viewpoint event detection from social media stream relates to the discovery and filtering of UGC. Twitter has recently made advances in this with its updated search, which incorporates aspects of event detection by using Amazon's Mechanical Turk service to detect and verify trending topics and breaking news events by using human evaluators to categorise search queries and provide additional context. While the approach yields good results, it relies heavily on manual input from Mechanical Turk - an online crowdsourcing platform where the human workers perform the HIT (human intelligence tasks) that computers are unable to. Hence, a framework that reduces manual effort involved in the detection of the events would be a huge asset to the field of journalism.

A location-based event detection framework aims to address the "where" in the fundamental Big-5 information-gathering questions (who, what, where, when, why), depicted in Figure 2. Metadata such as GPS coordinates, user-specified location information (e.g., from Twitter, Facebook profiles), and even more advanced methods such as landmark detection from image and video data can all be exploited to associate the location of UGC in the event detection process. In this paper we focus on Twitter, where four types of location information are considered most relevant for detecting the location of an event as follows:

- *Geo-tagged tweets*: These are tweets which are tagged with GPS coordinates. These are the most straightforward to process for location identification. However, only a small fraction of tweets (~1%) include GPS coordinates [12].
- *User specified profile location*: These are the information that a user presents in his/her profile information, normally identifying their residential location. Studies show roughly 3% of Twitter users include location in their profiles

[14]. Time zone information can also provide an indication of where a user is located.

- *Entity extraction and NLP techniques*: These use entity extraction and natural language processing techniques for identification of 'place' type entities in microblog text.
- *Social network analysis*: This method leverages a user's social relationships and the spatial distribution of locations in his/her network for identification of potential locations [12].

Further in this section, different approaches to detect the location from the types of available location information (as mentioned above) are briefly explained and reviewed:

Geotagged Data and User Profile

User-generated content may contain geographic coordinates in its meta-data. Sakaki et al. [24] used Twitter to specifically detect an earthquake. They relied on the GPS location attached to tweets as well as user profile locations, both of which are in latitudinal and longitudinal format. Location approximation techniques are deployed using Kalman filtering (a Bayes filter variant that uses a set of signals to determine an estimation which is much more precise than single observation). The derived location is then queried using the Google Maps API to check the location on the map and establish the location of the user.

Fig. 2. The Big-5 information-gathering questions

Unankard et al. [30] followed a similar approach to extract event locations from tweets. The authors assigned a hotspot location to event clusters if sufficient correlation is found between the detected location and the event. Before calculating the correlation score they extracted the event location and user location. To find the user location, they depend on the geographic tags of the tweets and if that is not

available they use the user profile location. The geographic tags are derived from coordinates logged by the devices of the users (e.g., smart phones), and profile location information is derived directly from profile text. The geographic tag data is then queried using Google API to find the location name, and in the case of user-profile location they queried gazetteer database- a list of locations downloaded from GeoNames[1] and stored in a local database. For event location, they followed an NLP-based approach to extract location entities from the text. Once both types of locations are determined, a correlation score is calculated for the event using both the locations, which is then assigned to the cluster.

Entity Extraction and NLP Techniques

NLP techniques assist in observing the events, sentiments, and to extract information such as variety of entities and tagging them. In order to extract the location for an event from the user-generated content, the text data is processed through NLP tools to determine the entities and their context with respect to parts of speech (POS). Ritter et al. [19] tested the performance of Stanford Named Entity Recognition tagger [9] named entity taggers[2], and parts of speech taggers[3] on Twitter streaming data. Unankard et al. [30], discussed earlier, also applied NLP techniques to determine the event location along with user location. To extract the event location they processed the textual data of the tweets with Named Entity Recognition (NER) for which they used the Standard Named Entity Recogniser to identify the location entities from the text messages. The most frequent location in each cluster of the detected event gets assigned as the location of the event. A correlation score is calculated between event location and user location (discussed above), by computing the level of granularity each location derives (whether it derives a country, state, city, or place name).

User Social Network Analysis

A user's social network plays an important role in determining the user's location. Often when the content-based approaches (geo-tagged data, user profile location) fail to determine the location of a user, it is the user's social network that can help in understanding from where the user is posting the content. Sadilek et al. [23], created a location prediction system, named Flap, which implements a probabilistic model of human mobility and generates a graph of people's fine grained location based on their friendship graph. First they recreate the friendship graph based on content similarity between users and redefine the edges. The location is predicted on a dynamic Bayesian network of the user and friends (from recreated friendship graph). The input sequence consists of locations visited by a user's friends (during supervised learning, the user's location was also given as input).

Jurgens [12] proposed a method based on a combination of spatial label propagation and a final location selection method to infer the geographical location of a user. In the same line, it is shown that an individual's location prediction is accessible to social network providers [2], where a relation is mapped between user's geographic locations and the friendship relationships on Facebook network. The same

[1] http://www.geonames.org/
[2] http://nlp.stanford.edu/software/segmenter.shtml
[3] http://nlp.stanford.edu/software/tagger.shtml

mapped relation is then reverted to infer a user's probable location based on his/her social relationships. Location-based social networks (LBSN) materialise directly the combination of social and geographical proximity and their study provide insight into how location and proximity impact social relationships [26].

The next section introduces the proposed framework that leverages the aforementioned techniques for inferring the location of an event.

3 A Framework for Location-Based Event Detection from UGC

Our framework for event detection focuses on detecting events from streams of disparate social media sources. Briefly, it aims to detect events as they happen from a stream of various social media modalities. We distinguish three types of events which require different methods:

- Breaking news - Events that are current and that were not precipitated or known a priori, e.g., a plane crash or the death of a prominent figure. Detecting breaking news from UGC requires real-time stream processing and analysis of popular social media sources.
- Running stories - This requires ongoing analyses of previously breaking news events or ongoing coverage of scheduled events.
- Scheduled events - These have known start and end dates (e.g., the Olympic Games) and can be followed by selecting certain topics or following users influential in those topics. However, unpredictable sub-events within these must be automatically detected in ways similar to breaking news.

A straightforward approach (disregarding the problem of data access), would consist of a stream processor that extracts named entities and hashtags from streaming UGC and monitors the frequency acceleration of these terms over time. The aim would be to flag the terms accelerating most frequently over time, as well as particular named entities, as potentially interesting events for journalists. Alternatively, first story detection (FSD) could be adapted to a more general framework for event detection from social media for breaking news, by treating tweets or other micro blog entries in a similar fashion. Being able to identify when something is happening in relation to a particular topic or location can be used as a first step for discovery, using entities detected through NLP.

There are other notable gaps in the literature of event detection. These include seamless interoperability between data sources and modeling different viewpoints of certain events, not only via different information streams. We propose a framework that will advance the state of the art by developing a linked data-based approach to event detection. This approach will build on previous work, e.g., [11], to develop a scalable and linked framework for flexible event detection from UGC streams. With respect to event detection in the information extraction context, recent efforts based on the work of Ritter [20] indicate that unsupervised approaches leveraging unlabeled data are promising. However, to our knowledge, NLP based event extraction for social media approaches do not leverage semantics. We envisage a framework that uses ontology-based IE expertise and improves on such semi-supervised approaches by exploiting existing linked data resources as well as available event ontologies.

Figure 3 depicts the proposed framework with two variations depending on where the location information is desired. Scenario A uses location information pre-specified by the journalist as a filter, thus it would discover events in that specified location. Scenario B clusters all detected events based on location information extracted from the media, leaving the journalist free to select among clusters. Scenario A is better suited to events where prior knowledge about the location is known, for example running stories or scheduled events as described above. In the case of general breaking news events, it is likely that any filter would be too restrictive unless such an event is expected at the specified location.

Scenario B, on the other hand, can be expected to detect a wider range of events, and is suitable for any of the three event types, with the caveat of a higher processing requirement and leaving the journalist with more manual intervention. In Scenario B, as shown in figure 4, the input stream is processed through various location detection techniques (as discussed above - geo tagged tweets, NLP techniques using Stanford NER or third party library AlchemyAPI, or social network analysis of the user), which results in cluster formation of various locations. However, there is one aspect that is not covered in the above techniques and that is leveraging the Semantic Web to refine the location cluster formation. Thus, this framework proposes using location knowledge relation mapping (from Linked Geo Data[4] or OpenGeoSpatial[5] data) to harvest the geographical proximities (two places from same country or from same city) so as to refine our location clusters, which means that if there are two different tweets: "*bombing at Boylston Street*" and "*Explosion at Boston marathon near finish line*", we are likely to infer that *Boylston Street* is also conveying information about *Boston* or on a more macro level - *United States*, apparently both are conveying information about explosions in United States. However, this might be a computationally time costly process as it may call for querying the location knowledge data (graphs) each time a new text document is processed.

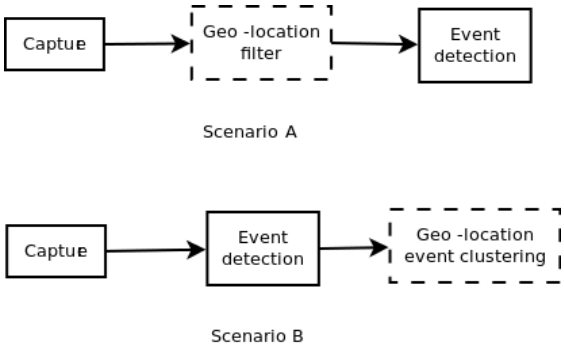

Fig. 3. Location-based event detection frameworks in two scenarios

[4] http://linkedgeodata.org/About
[5] http://www.opengeospatial.org/standards/geosparql

Once the location clusters are formed, the system (owner) has the leverage to mine and process the text data corresponding to each location cluster through event detection techniques. This can be achieved by breaking the input stream (of location cluster) into small time segments and analysing the *burst* keywords (meanwhile filtering out stop words) in the timeframes. The keywords whose frequency exceeds a *threshold* are considered and their corresponding text content (tweets) is taken into account. Further, *tweets* are matched against each other through vector cosine similarity (cosine similarity between text documents - converts text as vectors and evaluates the distance between vectors thus inferring the degree of relatedness between two text vectors) and if the matching output is more than a certain minimum value than they are cascaded into the same clusters, which signifies a particular *trend/event* against the *burst* keyword. This process is performed on incoming streaming data continuously and it results in clusters of events within the cluster of locations. That means, for any given location there are different clusters signifying different events.

Now that we have the location, event, and some named entities retrieved through an intermediate process (where the text was processed through Entity taggers), we can use those entities to explore and learn more about their co-occurrence significance by querying archived data from Linked Open Data (LOD) or other archival data. This would yield into a more insightful information about the event and generate more sense about its significance. Until this point, we have the events based on locations, related archival/historical data (archived stories relating to the entities of the current event), and users particular to events. What we do not have is a mechanism to verify the credibility and authenticity of the information.

In the next phase of the framework, we suggest ideas that can assist the journalists in determining the veracity. This is a multi-level process of filtering and reducing the tweets data from the clusters formed in the event detection process to a smaller set of tweets and its users, which can be further manually analysed. There are different approaches for determining the veracity of the content. Filtering the data based on multiple observations are likely to yield better results, for instance, merging the two different approaches - one based on a human centred design approach as proposed by Diakopoulos et al. [6] and other based on social network analysis.

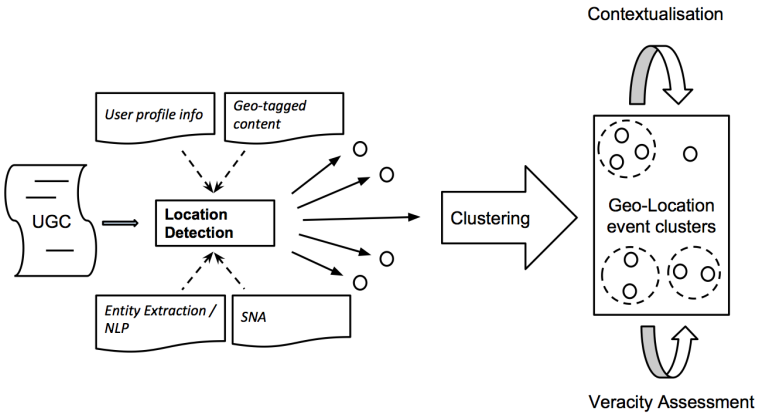

Fig. 4. Location-based event detection in Scenario B

A classifier is trained where tweets are analysed whether the user is a witness to the event he has posted or not. Text of the tweet is analysed against a dictionary[6], which contains a rich set of words across several language dimensions (for instance *affect*, *cognitive mechanism*, *health* etc.). It is hypothesised that such words are indicative of a user being witness and carrying an experiential authority pertaining to the event the user has posted. The classifier matches the tweets (which are not *re-tweets*) against the dictionary which may reflect the user's experience in time and space during an event. The tweets (with its user) which contain at least one such word from dictionary are marked as *eyewitness.* Now, all such users are further filtered through another check to narrow down to a more credible list of users. This can be achieved by analysing the social network/profile of the user, for example, we can guess users' location on the basis of geo-tagged information of the location in the profile, and also determine how reliable a profile is. Following the work by Castillo et al. [4], the *user-based* features in a data can be given a priority while assessing the authenticity of a user profile. The *user-based* features consider a user's characteristics based on the frequency of tweets, number of followers, and followee of the user. There are other features too, which can be considered while determining the trust value of the information, such as length of tweets, usage of special characters in text, number of re-tweets, and presence of a URL etc. A certain weight is given to the users who share the same location as that of the event, and even a higher weight is given to the users who have at least a threshold minimum number of followers, followees, and tweets (this threshold value can be based on a training data). Information propagated through credible users is seen to be as reliable, and the *user-based* features are indicative of users' reputation and hence credibility [4]. Once the user list and the corresponding tweets from an event are narrowed down to credible and authentic ones, the manual verification efforts can be greatly reduced.

The above described processes are likely to culminate into the following functionalities: *news search*, *popular/trending issues*, *related content (from archives)*, *news by location*, *real time trends,* and *verified news*. This would result in a platform for a panoramic view of the real time data of the user generated content over various social media platforms, and provide a leverage to perform analysis and visualize the history of an event or related events.

Framework Evaluation

The effectiveness of the proposed location-based event detection framework must be evaluated through experiments on real data. Each module can be evaluated separately to measure suitability in the framework as a whole. For example, the event detection module can be tested using manually annotated event streams. The accuracy of the location extraction component can be similarly tested and evaluated using manually annotated data much like the TREC-2009 Blog Track where a retroactive event detection (RED) approach was taken [15], i.e., event streams are captured, annotated, and then used to train, validate, and test event detection accuracy. Metrics for accuracy range from classification accuracy (%) to information retrieval metrics such as *precision* and *recall*. For example, the evaluation of the Twitter-based event detection framework in [33] used precision, a measure of the proportion of relevant

[6] http://liwc.net/

events detected from all events detected. More formally, $\text{precision} = \frac{|\text{relevant events} \cap \text{retrieved events}|}{|\text{retrieved events}|}$.

Because location predictions can be approximate, metrics used to evaluate the module should be sufficiently flexible to reward high-precision, high-accuracy results and penalise low-precision results. Evaluation of the framework as a whole, on the other hand, is envisioned be more end-user focused, with hands-on use by real journalists. This stage of evaluation can only be carried out when a working prototype has been implemented comprising the necessary modules.

4 Impact and Conclusions

We have outlined a general framework for location-based event detection of UGC for journalists. The framework is part of a larger vision of Social Semantic Journalism which aims to address the many technological challenges facing journalists today by aiding in the sourcing, aggregation, filtering and verification of UGC for news reportage; the larger goal being to reduce the workload facing journalists today.

The proposed framework illustrates two potential scenarios that can be targeted. The first uses pre-specified location information as a filter for an event where that location may be known *a priori*, and lends itself to running stories or scheduled events. The second clusters all detected events by location, ultimately resulting in more manual work on the part of the journalist to inspect and identify event clusters of interest. The aim of the framework is to provide a platform and a tool to assist journalists as well as users to generate insights over several dimensions of UGC analysis.

The proposed framework is an attempt to knit together various level techniques which have till now been, to the best of our knowledge, researched or existing as modular services rather than being a part of a large information retrieval system. It further relies on the Semantic Web technologies to refine its inferences about locations and insights about the news/topic from several layers of mappings in Linked Open Data. However, the challenges which will need to be addressed and countered are pertaining to the immense magnitude of noise that comes with the information.

References

1. Allan, J., Papka, R., Lavrenko, V.: On-line new event detection and tracking. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 37–45. ACM (1998)
2. Backstrom, L., Sun, E., Marlow, C.: Find me if you can: improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th International Conference on World Wide Web, pp. 61–70. ACM (2010)
3. Becker, H., Naaman, M., Gravano, L.: Beyond trending topics: Real-world event identification on Twitter. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, ICWSM 2011 (2011)
4. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web, pp. 675–684. ACM (2011)

5. Chen, Y., Amiri, H., Li, Z., Chua, T.S.: Emerging topic detection for organizations from microblogs. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 43–52. ACM (2013)
6. Diakopoulos, N., De Choudhury, M., Naaman, M.: Finding and assessing social media information sources in the context of journalism. In: Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems, pp. 2451–2460. ACM (2012)
7. DOMO, How Much Data is Created Every Minute? http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/ (retrieved October 16, 2013)
8. Elloumi, S., Jaoua, A., Ferjani, F., Semmar, N., Besançon, R., Jaam, J., Hammami, H.: General Learning Approach for Event Extraction: Case of Management Change event. Journal of Information Sciences (2012)
9. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 363–370. Association for Computational Linguistics (2005)
10. Heravi, B.R., McGinnis, J.: A Framework for Social Semantic Journalism. In: First International IFIP Working Conference on Value-Driven Social & Semantic Collective Intelligence (VaSCo), at ACM Web Science 2013, Paris, France (May 2013)
11. Hromic, H., Karnstedt, M., Wang, M., Hogan, A., Belák, V., Hayes, C.: Event Planning in a Stream of Big Data. In: LWA Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML). Workshop at LWA: Lernen, Wissen, Adaption (2012)
12. Jurgens, D.: That's What Friends are for: Inferring Location in Online Social Media Platforms Based on Social Relationships. In: Seventh International AAAI Conference on Weblogs and Social Media (2013)
13. Kijak, E., Gravier, G., Gros, P., Oisel, L., Bimbot, F.: HMM based structuring of tennis videos using visual and audio cues. In: Proceedings of the 2003 International Conference on Multimedia and Expo, ICME 2003, vol. 3, p. III-309. IEEE (2003)
14. Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., Shook, E.: Mapping the global Twitter heartbeat: The geography of Twitter. First Monday 18(5) (2013) (n. pag. Web. August 9, 2013)
15. Macdonald, C., Ounis, I., Soboroff, I.: Overview of the TREC 2007 Blog Track. In: TREC, vol. 7, pp. 31–43 (2007)
16. Petrovic, S., Osborne, M., Lavrenko, V.: Streaming first story detection with application to twitter. In: Proceedings of NAACL, vol. 10 (2010)
17. Qian, X., Liu, G., Wang, H., Li, Z., Wang, Z.: Soccer video event detection by fusing middle level visual semantics of an event clip. In: Qiu, G., Lam, K.M., Kiya, H., Xue, X.-Y., Kuo, C.-C.J., Lew, M.S. (eds.) PCM 2010, Part II. LNCS, vol. 6298, pp. 439–451. Springer, Heidelberg (2010)
18. Rahmanzadeh Heravi, B., Boran, M., Breslin, J.: Towards Social Semantic Journalism. In: Sixth International AAAI Conference on Weblogs and Social Media (2012)
19. Ritter, A., Clark, S., Etzioni, O.: Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1524–1534. Association for Computational Linguistics (2011)
20. Ritter, A., Etzioni Mausam, O., Clark, S.: Open domain event extraction from Twitter. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2012), pp. 1104–1112. ACM, New York (2012), doi:10.1145/2339530.2339704

21. Rosen, J.: Definition of Citizen Journalism (2008), http://www.youtube.com/watch?v=QcYSmRZuep4 (retrieved October 16, 2013)
22. Rui, Y., Gupta, A., Acero, A.: Automatically extracting highlights for TV baseball programs. In: Proceedings of the Eighth ACM International Conference on Multimedia, pp. 105–115. ACM (2000)
23. Sadilek, A., Kautz, H., Bigham, J.P.: Finding your friends and following them to where you are. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, pp. 723–732. ACM (2012)
24. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, pp. 851–860. ACM (2010)
25. Sayyadi, H., Hurst, M., Maykov, A.: Event detection and tracking in social streams. In: ICWSM 2009, pp. 311–314 (2009)
26. Scellato, S., Noulas, A., Lambiotte, R., Mascolo, C.: Socio-Spatial Properties of Online Location-Based Social Networks. In: ICWSM, vol. 11, pp. 329–336 (2011)
27. Shirazi, A., Rohs, M., Schleicher, R., Kratz, S., Müller, A., Schmidt, A.: Real-time nonverbal opinion sharing through mobile phones during sports events. In: Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems, pp. 307–310. ACM (2011)
28. Sonderman, J.: 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/ (retrieved October 16, 2013)
29. Statisticbrain, Twitter Statistics, http://www.statisticbrain.com/twitter-statistics/
30. Unankard, S., Li, X., Sharaf, M.A.: Location-Based Emerging Event Detection in Social Networks. In: Ishikawa, Y., Li, J., Wang, W., Zhang, R., Zhang, W. (eds.) APWeb 2013. LNCS, vol. 7808, pp. 280–291. Springer, Heidelberg (2013)
31. Van Rijsbergen, C.J.: Information Retrieval (1979) ISBN 0-408-70929-4
32. Weng, J., Lee, B.: Event detection in Twitter. In: Proc. of ICWSM 2011, pp. 401–408 (2011)
33. Weng, J., Lee, B.-S.: Event Detection in Twitter. In: ICWSM (2011)
34. Xu, C., Zhang, Y.F., Zhu, G., Rui, Y., Lu, H., Huang, Q.: Using webcast text for semantic event detection in broadcast sports video. IEEE Transactions on Multimedia 10(7), 1342–1355 (2008)
35. Xu, M., Maddage, N.C., Xu, C., Kankanhalli, M., Tian, Q.: Creating audio keywords for event detection in soccer video. In: Proceedings of the 2003 International Conference on Multimedia and Expo, ICME 2003, Vol. 2, pp. II-281. IEEE (2003)
36. YouTube statistics (2013), http://www.youtube.com/yt/press/statistics.html (retrieved October 16, 2013)