A Framework for Collecting, Extracting and Managing Event Identity Information from Textual Content in Social Media

Debanjan Mahata
Department of Information Science
University of Arkansas at Little Rock
Little Rock, Arkansas, USA

Social media has brought a paradigm shift in the way people communicate with each other. It has gone from being just a medium to a global medium of communication between people. Different types of social media platforms provide multiple venues to people for sharing first-hand experiences and exchange information about real-life events. It has become an indispensable source for disseminating news and real-time information about current events, using websites like Twitter, Facebook, Instagram, Flickr, Youtube, Vine, etc, that allow users to post short textual messages accompanied with images and videos. At the same time users also share their detailed citizen journalistic experiences in the form of diaries through different blogging platforms like Blogger, Wordpress, Medium, etc. Studies have shown the importance of different social media platforms as a news circulation service [**?**], and a source for gauging public interest and opinions [**?**, **?**, **?**, **?**]. It's efficacy as a real-time citizen-journalistic source of information has been recently harnessed in detection, extraction and analysis of real-life events [**?**, **?**, **?**]. The activities of users producing content in social media has also been studied for gaining deep insights about how they group together to form communities around topics related to real-life events [**?**, **?**, **?**], and lead to collective action [**?**, **?**].

With the popularity of social media there has been proliferation of unstructured textual information about different real-life events, in the Internet. The information gained by identifying and tracking social media content

expressing live reporting of an event, recent updates related to the event, insightful opinion about the different named entities (people, place, organization, etc) directly or indirectly involved with the event, summarization of content, among others, could prove to be extremely valuable for monitoring and gaining deeper actionable insights. There are tremendous applications in the areas of real-life event analysis, event management, opinion mining, reference tracking, online targeted marketing, recommendation engines, cyber security, enterprise data integration, among others. Thus, there is a need of a generic framework that has the following capabilities:

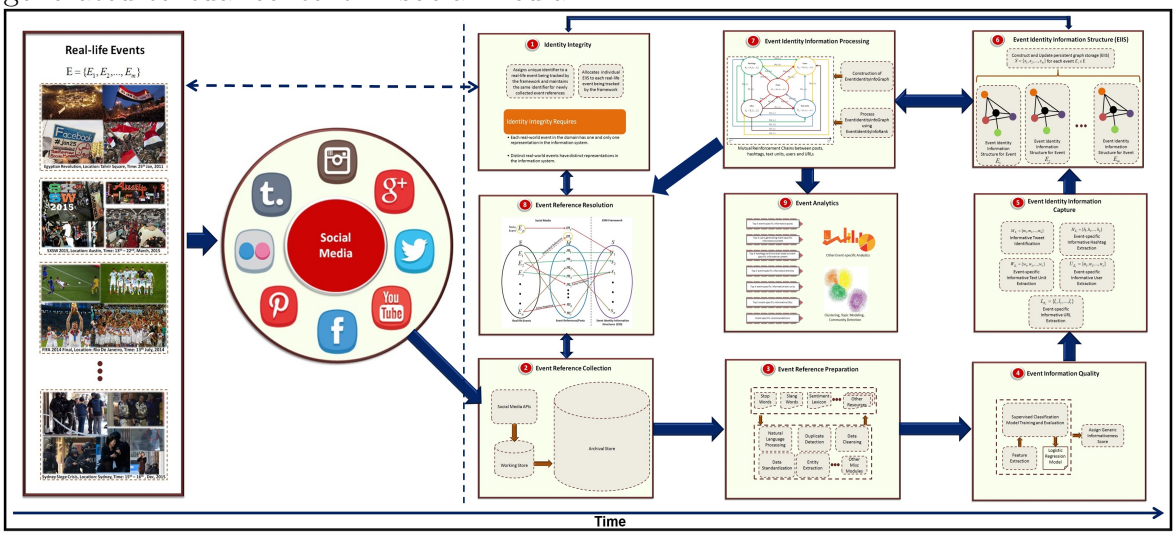
- can collect different types of textual content produced in social media related to an event
- extract information that acts as an identity of the event used for characterizing it
- maintain the extracted event identity information persistently for resolving constantly produced new content and discovering important event-specific information.

The problem of collecting and extracting event identity information from social media is very similar to the task of event detection and tracking from newswires [**?**, **?**]. However, in this thesis, we add new components of creating identity structures of an event and managing the tracked information persistently over time. In order to make our task well defined we avoid the task of detecting unidentified events, and instead track a pre-specified set of events. Also, the domain of social media poses additional challenges. News articles most often adhere to grammatical, syntactical and formal structures of writing, that are not common in the realm of social media. The user generated content in social media is most often colloquial, short, noisy and lack proper grammatical structures. This makes it a challenging task for the state-of-the-art natural language processing techniques to extract useful information and perform tasks like entity extraction and parts-of-speech tagging that lies at the core of the previous research on event detection and tracking.

The work presented in this thesis establishes the conceptual design and implementation of a framework capable of collecting, extracting and persistently managing event identity information from user generated textual content shared in social media (shown in Figure 1.1). The approach of the presented work is from the perspective of Entity Identity Information Management (EIIM) [**?**], with basic tenets of information quality at its core.

Towards this objective, different challenges of mining high quality information from social media text is discussed and a patent pending novel approach to tackle the challenges for identifying event-specific informative content is explained, which lies at the heart of the framework. It further explores the applications of the research and concludes by pointing to different future directions of the work.

Figure 1: Event Identity Information Management (EIIM) Life Cycle for user generated textual content in social media



Some of the main contributions of the work are:

- Extending the Entity Identity Information Management model [**?**] from the closed world domain of Master Data Management (MDM) to the open and unstructured domain of social media.
- Design and implementation of an *Event Identity Information Management* framework that is capable of tracking and identifying event-specific information from long as well as short user generated textual content in social media. Towards this objective a data processing pipeline named *Event Identity Information Management Life Cycle* is developed (Figure 1.1), which is capable of :
  - collecting event related real-time content generated in social media

  - pre-processing them using natural language processing techniques
  - identifying high quality informative sources of information
  - extracting event-specific information in order to create *Event Identity Information Structures* (EIIS) for persistently storing and characterizing the salient and high quality event related information
  - identifying event-specific informative content produced in social media

- Implementation of a supervised classifier in the domain of short and informal social media textual content, for segregating high quality informative messages having higher chances of containing event related information from the low quality non-informative ones.

- Analysis of informative and non-informative event related content from 3.8 million short textual social media messages.

- A novel model that leverages mutually reinforcing relationships between blog posts and named entities mentioned in them, and simultaneously ranks blogs as well as the named entities, allowing identification of event-specific content and further analysis of event-specific information.

- A novel model based on principle of mutual reinforcement that takes into account the semantics of relationships between short textual *social media messages*, *hashtags*, *text units*, *URLs* and *users*, and represent them in a graph structure - *EvenIdentitytInfoGraph*. A scalable graph processing iterative algorithm -*EventIdentityInfoRank*, is implemented for ranking the nodes of the *EventIdentityInfoGraph*. The algorithm is capable of simultaneously ranking *social media messages*, *hashtags*, *text units*, *URLs* and *users* in terms of event-specific informativeness providing deeper insights into the identity of an event.

- Evaluate the proposed techniques against popularly used baseline techniques using large scale datasets.

Already published work as well as upcoming publications that represents our contributions related to specific topics covered by the broad area of research as presented in this thesis are given below.

Related Filed Patent

- A System for Collecting, Ranking and Managing Entity Identity Information from Social Media (US 62135258). Inventors: **Debanjan Mahata** and John R. Talburt, Assignee: The Board Of Trustees Of The University Of Arkansas.

Related Award

- **Debanjan Mahata** and John R. Talburt. *Chatter that Matter : A Framework for Collecting, Extracting, and Managing Event Identity Information from Short Social Media Text.* Student Research and Creative Works Expo, Graduate Competition, University of Arkansas at Little Rock, April, 2015. (Awarded First Place in Engineering and Information Technology).

Related Publications

- **Debanjan Mahata**, John R. Talburt and Vivek Kumar Singh; *Identifying and Ranking of Event-specific Entity-centric Informative Content from Twitter.* 20^{th} International Conference On Applications Of Natural Language To Information Systems (NLDB 2015), Passau, Germany. $17^{th} - 19^{th}$ June, 2015.

- **Debanjan Mahata** and John R. Talburt; *A Framework for Collecting and Managing Entity Identity Information from Social Media.* 19^{th} International Conference on Information Quality, Xi'An, China.

- **Debanjan Mahata** and Nitin Agarwal; *Identifying Event-specific Sources from Social Media.* Online Social Media Analysis and Visualization. Lecture Notes in Social Networks, Springer, Kawash, Jalal (Ed). January, 2015.

- Nitin Agarwal, **Debanjan Mahata**, and Huan Liu. *Time-and Event-Driven Modeling of Blogger Influence.* Encyclopedia of Social Network Analysis and Mining. Springer New York, 2014. 2154-2165.

- **Debanjan Mahata** and Nitin Agarwal. *Learning from the crowd: An Evolutionary Mutual Reinforcement Model for Analyzing Events.* Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. IEEE, 2013.

- Nitin Agarwal, and **Debanjan Mahata**. *Grouping the Similar among the Disconnected Bloggers*. Social Media Mining and Social Network Analysis: Emerging Research (2013), 54.

- **Debanjan Mahata**, and Nitin Agarwal. *What does everybody know? identifying event-specific sources from social media*. IEEE Fourth International Conference on Computational Aspects of Social Networks (CASoN), 2012.

- **Debanjan Mahata** and Nitin Agarwal. *Analyzing Event-specific Socio-Technical Behaviors Through the Lens of Social Media*. The International Sunbelt Social Network Conference (Sunbelt XXXII) organized by the International Network for Social Network Analysis (INSNA), March 12-18, 2012, Redondo Beach, California.

- Vivek Kumar Singh, **Debanjan Mahata**, and Rakesh Adhikari. *Mining the blogosphere from a socio-political perspective*. IEEE International Conference on Computer Information Systems and Industrial Management Applications (CISIM), 2010.

- Vivek Kumar Singh, Rakesh Adhikari, and **Debanjan Mahata**. *A clustering and opinion mining approach to socio-political analysis of the blogosphere*. IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2010.

Related Submitted Publications

- **Debanjan Mahata**, John R. Talburt, Vivek Kumar Singh and Rajesh Piryani; *Chatter that Matter: A Framework for Identifying and Ranking Event-specific Informative Tweets*. 18^{th} International Conference on Text, Speech and Dialogue, Plzen, Czech Republic (Notification Due: May 10, 2015)

- **Debanjan Mahata**, John R. Talburt and Vivek Kumar Singh; *A Framework for Collecting, Extracting and Managing Event Identity Information from Twitter*. 20^{th} International Conference on Information Quality, M.I.T, Boston (Notification Due: April 30, 2015)

- **Debanjan Mahata**, John R. Talburt and Vivek Kumar Singh; *From Chirps to Whistles : Discovering Event-specific Informative Content*

from Twitter. Proceedings of the 7^{th} Annual ACM Web Science Conference. ACM, 2015, Oxford, England (Notification Due: April 30, 2015)