

Understanding Events Through Analysis Of Social Media

Amit Sheth

Kno.e.sis, Dept. of Computer Science
and Engineering

Wright State University, Dayton, OH

amit@knoesis.org

Hemant Purohit

Kno.e.sis, Dept. of Computer Science
and Engineering

Wright State University, Dayton, OH

hemant@knoesis.org

Ashutosh Jadhav

Kno.e.sis, Dept. of Computer Science
and Engineering

Wright State University, Dayton, OH

ashutosh@knoesis.org

Pavan Kapanipathi

Kno.e.sis, Dept. of Computer Science
and Engineering

Wright State University, Dayton, OH

pavan@knoesis.org

Lu Chen

Kno.e.sis, Dept. of Computer Science
and Engineering

Wright State University, Dayton, OH

chenlu@knoesis.org

ABSTRACT

Users are sharing vast amounts of social data through social networking platforms accessible by Web and increasingly via mobile devices. This opens an exciting opportunity to extract social perceptions as well as obtain insights relevant to events around us. We discuss the significant need and opportunity for analyzing event-centric user generated content on social networks, present some of the technical challenges and our approach to address them. This includes aggregating social data related to events of interest, along with Web resources (news, Wikipedia pages, multimedia) related to an event of interest, and supporting analysis along spatial, temporal, thematic, and sentiment dimensions. The system is unique in its support for user generated content in developed countries where Twitter is popular, as well as in support for SMS that is popular in emerging regions.

Categories and Subject Descriptors

D.3.3 [Social Networks]: Social computing systems, Spatio-Temporal-Thematic analysis

General Terms

social networks, social media, situational awareness, events

Keywords

Twitris+, Twitris, Semantic Social web, spatio-temporal-thematic-sentiment analysis, tweet, SMS, user generated content

1. Need and Opportunity: Event centric data on social media and its analysis

Social media is growing in an unprecedented way. Twitter's

100+ million users send out over 65 million tweets a day. In the recent years, we have seen Twitter as a significant platform for disseminating information about and understanding the evolutions of significant events of local and global importance. Global events where Twitter played pivotal roles include Mumbai Terrorist Attack, Iran Elections, Haiti Earthquake, and US Healthcare debate. Equally important are events of local and regional importance such as Gilroy garlic festival and Ohio State Fair that reflect traditional values of the culture. While Twitter has captured the imagination of developed world, there is an equally impressive revolution going on in the developing world and emerging regions. Perhaps the most celebrated example is that of Ushahidi¹ which was started by gathering testimonies to map reports of violence in Kenya after the post-election fallout at the beginning of 2008. Since then it has been used for many crisis management applications. Ushahidi and a number of other platforms such as eMoksha², Kiirti³ and SMS GupShup⁴ provide an alternative of a much larger segment of 4.1 billion users today who use mobile phones that are not smartphones. These sites and related applications support exchange and sharing of user generated content (UGC) using SMS. The range of applications include emergency monitoring and management (e.g., pakreport⁵), social activism (e.g., coalition against corruption⁶), citizen awareness and engagement for strengthening democracies, education, rural development and public health. Many of interactions over these platforms are centered on events and activities. In some events such as Haiti earthquake, both Twitter and SMS based UGC was extensively used.

While current technologies have enabled easy access and sharing of social media content, there is a serious need to aggregate relevant social media and web content, and analyze them to understand events as they unfold. For example, the following are

¹ ushahidi.org

² eMoksha.org

³ Kiirti.org

⁴ smsgupshup.com

⁵ pakreport.org

⁶ pac-cac.kiirti.org

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WWW'11, March 28–April 1, 2011, Hyderabad, AP, India.

Copyright 2010 ACM 1-58113-000-0/00/0010...\$10.00.

some interesting questions that could have been asked during some recent events for improving situational awareness:

- During the Mumbai Terrorist Attack, what are the main topics (key phrases) that were discussed in the tweets originating from Mumbai during each of the three days of carnage?
- During Haiti earthquakes or recent Pakistan floods, what are the primary immediate requirements in rescue situations, where are the possible locations from which supplies that match requirements are offered?
- During recent Malaria outbreak in Rampir No Tekra (Gujarat’s largest slum near Gandhi’s Sabarmati Ashram), identify early SMSs and/or Tweets that mentioned water logging, and plot them if possibly by time and location.

Current keyword based search or other mechanisms (e.g., hashtags) are simply too inadequate for such purposes. There has been initiative in this regard in Tweak-The-Tweet⁷ project. The Twitris+ system which we present provide significantly advanced capabilities for finding answers to the questions similar to those discussed above. Ongoing work by us and others also pursue methodological studies that will lead to understanding of how to identify various facets of an event or an evolving situation (e.g., rescue versus recovery phases of a crisis), how to promote interactions and effective coordination among participants whether individual or institutional (e.g., Red Cross), etc. Section 2 describes challenges and proposed approach for providing solutions to this problem space. Section 3 describes our proposed system **Twitris+**, which is an extension of Twitris [2] to support SMS data and additional features.

2. Key technical challenges and proposed approach

There are a large number of technical challenges for developing solutions. The first and relatively easy challenge is that of finding and aggregating event centric data from different modes that encompass Web, microblogs and SMS sources. This is generally done using APIs provided by the systems (e.g., APIs from Twitter and Ushahidi). The second and tougher challenge relates to informal text analysis. Microblogs and SMS data represent several challenges for text processing and analysis [2,3,5]. Casual text form of the content necessitates the need to go beyond conventional text processing approaches. Moreover,

mumbai	1.4553	pakistan pres promised	1.0065
photographers capture images	1.3998	mumbai attacks	0.9594
images of mumbai	1.2792	foreign relations	0.9490
foreign relations perspective	1.2165	rejected evidence	0.8741
attacks in mumbai	1.1261	evidence provided	0.8741
photographers capture	1.0986	uk indicating	0.8741
capture images	1.0986	mumbai attacks in	0.7927
india prime minister	1.0839	rejected evidence provided	0.7916
country of india	1.0280		

Event descriptors sorted by their TFIDF scores

presence of conventions on these social media platforms such as mentions (denoted by @), shortened URL resources, user names, hashtags etc requires us to preserve their semantics while identifying and processing them. Also, conversational practices such as message forwarding like retweeting on twitter and mentions tend to create a statistically significant bias in the corpus due to repetition of the text. To perform statistical computations such as TFIDF computation on a changing corpus, requires fundamental changes in the way these computations are defined and performed.

Often used text processing techniques such as those involving parsing or other NLP techniques that work reasonably well on regular text do not work well on informal text. The use of abbreviations, improper grammar, and lack of significant context complicate the problem. Correspondingly, we developed enhanced statistical and learning techniques that utilize background knowledge [2,3,6] and improved text processing with spatio-temporal-thematic (STT) bias resulting in significantly better results. Figure-1 shows one such example for STT bias. To improve understanding of short and low-context text and enhance its analysis, we also exploit relevant and complementary data from other Web sources, such as news and other Web resources pointed by embedded links including pictures from Twitpics and video from Youtube, as well as Wikipedia articles on the topic, if available. Furthermore, there is an increasing amount of metadata collected and supplied by the devices on which users create content, and a variety of approaches are needed to extract and normalize such metadata. The extracted metadata involves structured embedded metadata like spatial, temporal, thematic, profile metadata or annotations, network structure features and metadata from structured components of data like attention metadata (likes, views, listens etc).

3. Demonstration System: Brief description and a subset of capabilities

3.1 System overview

Twitris+ system, which we demonstrate here is accessible at <http://twitris.knoesis.org>. Current system can demonstrate spatial, temporal, thematic and sentiment analysis over a number of events by analyzing user posts (tweets or SMS). At the time of this writing, Twitris+ statistics, not accounting for SMS data, are given in Table-2. Data for new events and existing progressive events are continuously added, and significant scaling of the

foreign relations perspective	1.7185	photographers capture images	1.3028
india prime minister	1.5853	rejected evidence provided	1.2933
country of india	1.5295	mumbai attacks	1.2048
pakistan pres promised	1.5080	images of mumbai	1.1822
foreign relations	1.4510	mumbai	1.1083
rejected evidence	1.3758	mumbai attacks in	1.0797
evidence provided	1.3758	photographers capture	1.0017
uk indicating	1.3758	capture images	1.0017
attacks in mumbai	1.3293		

Event descriptors sorted by their enhanced spatio-temporal-thematic scores

Figure 1. Effect of STT bias for extracted event descriptors in ‘Mumbai Terror Attack 2008’: Descriptors generic to other spatial and temporal settings (e.g., mumbai and mumbai attacks) get weighted lower, allowing the more interesting ones to surface higher (e.g., foreign relations).

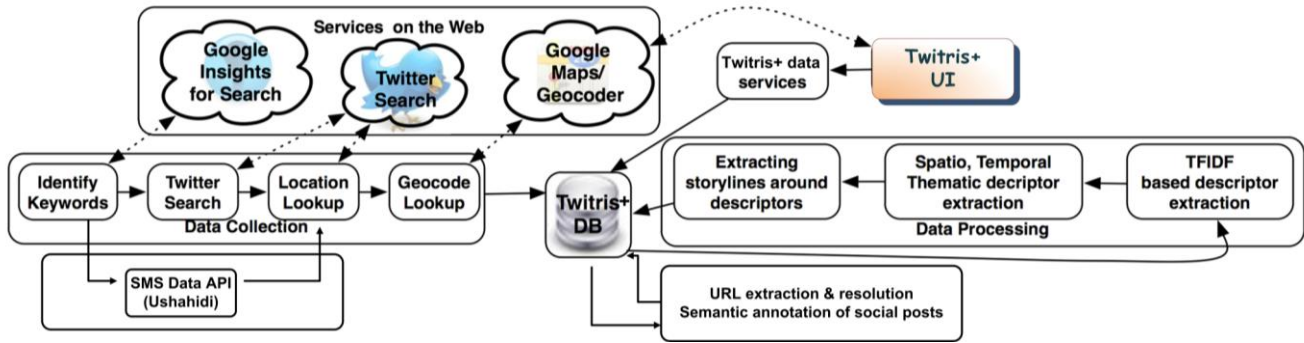


Figure 2. Twitris+ architecture: Data Collection & analysis

system is in process.

Table 1. Statistics of the collected data (by Oct. 2010)

Total number of extracted tweets	17.5 million
Processed Tweets	8 million
Cached unique location geocodes	595,301
Cached author locations	2.4 million
Extracted event descriptors	3.7 million
Extracted DBpedia entities	1.3 million
Extracted external URLs	649,165

Twitris+ is developed as a multi-layered system where each component acts as part of a pipeline. Figure-2 shows system architecture and Figure-3 shows functional overview of Twitris+ system.

Twitris+ takes a Semantic Social Web approach to detect social signals by analyzing massive, event-centric data through

- Analysis of casual text with spatio-temporal-thematic (STT) bias, to extract event descriptors.
- Capturing semantics from three contexts: internal context (context obtained by analyzing directly mentioned content, annotated entities and related posts based on event descriptors), external context (obtained from external sources by semantically following the theme of the current post) and mined internal context (entity-relationship, Sentiment Analysis).
- Use of deep semantics (using automatically created domain models [7]) to understand the meaning of standard event descriptors.
- Use of shallow semantics (semantically annotated entities) for knowledge discovery and representation.
- Semantic Integration of multiple external Web resources (news, articles, images and videos) utilizing the semantic similarity between contexts.

3.2 Extraction of Event Descriptors

Twitris+ processing starts with social signals collection; we have implemented a near real-time extractor to fetch topically relevant tweets using Twitter Search API. We have also integrated Ushahidi API for collecting SMS data. The volume of messages on a popular topic coupled with the short nature (140 characters for tweets) poses significant challenges for extraction and processing. We discuss key innovations in Section 2 that allow us to address these challenges.

Data collection is followed by a three step processing for finding N-gram summaries from tweets. First, it creates the Spatio-Temporal clusters of the tweet corpus surrounding an event. TFIDF computation is performed to fetch the n-grams from this set. Second step involves the association of spatial, temporal and thematic bias to these n-grams by means of enhancing the weights, while preserving the contextual relevance of these event descriptors. Finally, we create domain models automatically considering the event context and prominent event descriptors using Doozer [7]. These domain models are used to facilitate fine grained browsing of concepts and as an evidence to calculate the weight of extracted terms. We enhance the weights of descriptors that share a relationship with one or more terms from the semantic keyword cluster. Further details of the text processing algorithm are available in [2].

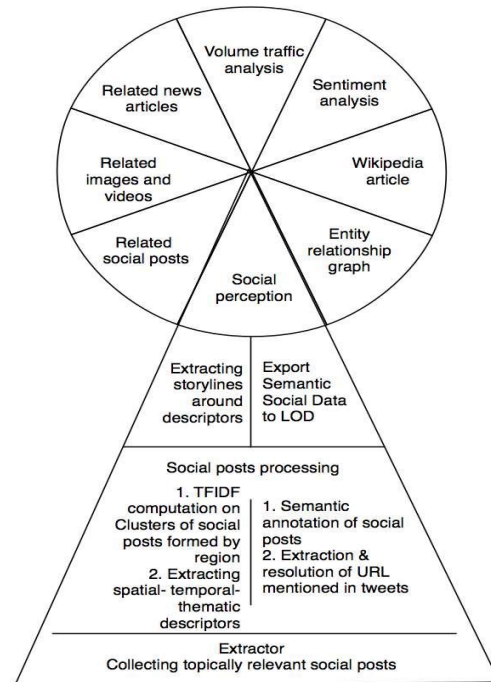


Figure 3. Twitris+ functional overview

Extracted event descriptors are used for finding implicit features of the content like sentiments and intentions expressed in the posts. These descriptors help users to quickly understand an event.

3.3 Twitris+ user interface

Twitris+ integrates the results of the data analysis (extracted descriptors and surrounding discussions) with emerging visualization paradigms to facilitate sensemaking. The current user interface facilitates effective browsing of the ‘when’, ‘where’, and ‘what’ slices of social perceptions behind an event and includes UI components to illustrate the theme, time and space. The UI also includes widgets for media (Related posts, News and Referenced articles, Wikipedia articles), entity-relationship graph, images, videos, traffic volume and sentiments. These widgets are dynamically populated depending on the user’s selections.

We will also demonstrate a number of additional features not described above such as:

- Real-time data analysis using Twarql [4] technologies, for annotations and management of streaming social data and representation of information from microblog posts and SMS as Linked Open Data⁸ for collectively analyzing social data for sensemaking.
- Dynamics of information diffusion on Twitter [5], showing a strong three-dimensional dynamic at play – the people involved (passionate advocate or an objective observer), the content being shared (fact-sharing or emotionally charged) and the connections between the people. Understanding these micro-level variables and their interactions will shed light on macro-level consequences e.g., political decisions, consumer behaviors, decision making in crisis management etc.
- Sentiment analysis for social signals to show different opinions from spatio-temporal aspects and thus provide a lens to understand the changes in the voice of the people. It can be an effective medium for understanding how people react to a certain event such as the US Healthcare debate and the Iran Presidential Election.

4. Conclusion

We discussed the need and opportunity for understanding event centric data on social media, followed by key challenges in addressing this problem space. Growing use of mobile phones has been significant source of connecting people in developing nations, whereas platforms like Twitter & Facebook has proven their potential to reach the mass in developed nations. We demonstrated a robust approach for analyzing social signals by analysis of aggregated social data shared on the Web in various forms with emphasis on short messages like microblog posts and SMS. We use background knowledge supported statistical and learning techniques for understanding informal text and finding associated metadata. Our extracted event descriptors with spatio-temporal and cultural bias give improved understanding of events of different scales and nature.

5. Acknowledgments

We acknowledge Pramod Ananthram, Vinh Nguyen, Ajith Ranabahu, Pablo Mendes, Alan Gary Smith, Michael Cooney and

Meenakshi Nagarajan (alumini) at Kno.e.sis Center, Wright State University for giving fruitful feedback and helping us build the system. This research was supported by NSF Award#IIS-0842129, titled "III-SGER: Spatio-Temporal-Thematic Queries of Semantic Web Data".

6. References

- [1] Sheth, A. 2009. *Citizen Sensing, Social Signals, and Enriching Human Experience*, IEEE Internet Computing, pp. 80-85..
- [2] Nagarajan M., Gomadam K., Sheth A., Ranabahu A., Mutharaju R. and Jadhav A. 2009. *Spatio-Temporal-Thematic Analysis of Citizen-Sensor Data - Challenges and Experiences*. 10th Intl Conf on Web Information Systems Engineering. (Oct 5-7, 2009). pp. 539 - 553..
- [3] Nagarajan M., Gruhl D., Pieper J., Robson C., Sheth A. 2009. *Entity Spotting in Informal Text*. The Semantic Web - ISWC 2009, Proceedings of 8th International Semantic Web Conference (Chantilly, VA, USA, October 25-29, 2009). pp. 260-276.
- [4] Mendes P., Passant A., Kapanipathi P., Sheth A. 2010. *Linked Open Social Signals*. IEEE/WIC/ACM International Conference on Web Intelligence (WI-10, Toronto, Canada, Aug. 31-Sep. 3, 2010).
- [5] Nagarajan M., Purohit H., Sheth A. 2010. *A Qualitative Examination of Topical Tweet and Retweet Practices*. 4th Int'l AAAI Conference on Weblogs and Social Media, ICWSM 2010, 295-298
- [6] Gruhl D., Nagarajan M., Pieper J., Robson C., Sheth A. 2010. *Multimodal Social Intelligence in a Real-Time Dashboard System*. Special issue on 'Data Management and Mining for Social Networks and Social Media', the VLDB Journal.
- [7] Sheth A., Thomas C. and Mehra P. 2010. *Continuous Semantics to Analyze Real-Time Data*. IEEE Internet Computing, vol. 14, no. 6, (Nov/Dec. 2010). pp. 85-89.
- [8] Nagarajan M. 2010. *Understanding User-Generated Content on Social Media*. Ph.D. Dissertation. Kno.e.sis Center, Wright State University.

⁸ <http://linkeddata.org>