UNIVERSITY OF ARKANSAS AT LITTLE ROCK

DOCTORAL THESIS

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

Author:
Debanjan Mahata

Supervisor:
Dr. John R. Talburt

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in

Integrated Computing
Information Quality Track
Department of Information Science

April 2015

Declaration of Authorship

I, Debanjan Mahata, declare that this thesis titled, 'A Framework for Collecting, Extracting and Managing Event Identity Information from Short Social Media Text' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ____

Date: ____

“Torture the data, and it will confess to anything.”

Ronald Coase, Economics, Nobel Prize Laureate

Abstract

With the popularity of social media platforms like Facebook, Twitter, Google Plus, etc, there has been voluminous growth in the digital footprints of real-life events in the Internet. The user generated colloquial and concise textual content related to different types of real-life events, produced in these websites, acts as a hotbed for researchers and organizations for extracting valuable and meaningful information. There has been significant improvement in natural language processing techniques for mining formal and long textual content often found in blogs and newspaper articles. But, it is still a challenging task to mine textual information from the social media channels producing terse, informal and noisy text with an unusual structure. For an event of interest it is necessary to detect and store event-specific signals from the noisy social media channels that allows to distinctively identify that event among all others and characterizes it for drawing actionable insights. These event-specific cues also forms its identity in the unstructured domain of social media. This identity information when mined and analyzed in a timely manner has tremendous applications in the areas of real-life event analysis, opinion mining, reference tracking, recommendation engines, cyber security, event management, among others. Thus, there is a need of a generic framework that can collect short textual content related to real-life events, extract information from them and maintain the information persistently for performing data analytics tasks, and tracking newly produced content as an event evolves. The patent pending work presented in this thesis establishes the design and implementation of an extendable framework enabling collecting, extracting and persistently managing identity information of real-life events from short textual content produced in social media. Towards this objective a pipeline of data processing components going through repeated processing cycles - *Event Identity Information Management Life Cyle* (EIIM) is proposed. A novel persistent graph data structure - *EventIdentityInfoGraph* representing the identity information structure of an event is implemented that forms the core component of the EIIM cycle. Mutually reinforcing relationships between event-specific social media posts, hashtags, text units, URLs and users, forming the vertices of the graph and denoting *event identity information units*, are defined and quantified. An iterative and scalable algorithm - *EventIdentityInfoRank* is proposed that processes the vertices of the graph and ranks them in terms of event-specific informativeness by leveraging the mutually reinforcing relationships. The ranked *event identity information units* are further used in tracking new event related content and extracting valuable event-specific information. Different components of the framework are tested and validated for real-time event related content generated in social media. The work is concluded by discussing about its novel contributions, practical applications in various other domains and envisaging future directions.

Acknowledgements

I would like to express the deepest appreciation to my committee chair Dr. John R. Talburt, who has shown the attitude and the substance of a genius. He continiously and persuasively conveyed a spirit of adventure in regard to research and scholarship, and an excitement in regard to directing innovation towards practical problems. Without his supervision and constant support this dissertation would not have been possible.

I would like to thank my committee members, Dr. Elizabeth Pierce, Dr. Ningning Wu, Dr. Russel Bruhn and Dr. Mathias Brochhausen, whose high quality contributions in the field of Information Science and Information Quality have inspired me to set high standards in my work, and kept me motivated. I would specially thank Dr. Mathias Brochhausen for devoting his valuable time for discussing about possible applications of ontologies in representing real-life events and the related information content in social media. I strongly consider it as one of the future directions of my research.

In addition, I thank Dr. Vivek Kumar Singh and his team from Banaras Hindu University, India, for collaborating with me and helping me to execute the necessary evaluation tasks in an unbiased way, including manual annotations and feedback. I also acknowledge the support of Mr. Jeff Stinson and Ms. Glediana Rexha for financially supporting the major part of my PhD by allowing me to work as a Graduate Assistant at TechLaunch, University of Arkansas at Little Rock.

I am extremely thankful to Dr. Abhijit Bhattacharyya (Associated Dean, Donaghey College of Engineering and Information Technology), for providing me with advise and encouragement from time to time. This acknowledgement page would be incomplete without thanking the immense support of my friends and family. I thank my parents, wife and friends (specially Pathikrit Bhattacharya, Subhashish Duttachowdhury and Meenakshisundaram Balasubramaniam) for not only their support but for their constant interest in my work and the discussions that I had with them. The conversations with them helped me to understand the information seeking behavior of various people from social media, with different perspectives.

Lastly, I thank University of Arkansas for providing me with the facilities, funds and a congenial environment for working towards my goal of PhD. I also acknowledge the Board Of Trustees Of The University Of Arkansas for filing a provisional patent of my work and encouraging me to pursue a path of innovation.

Contents

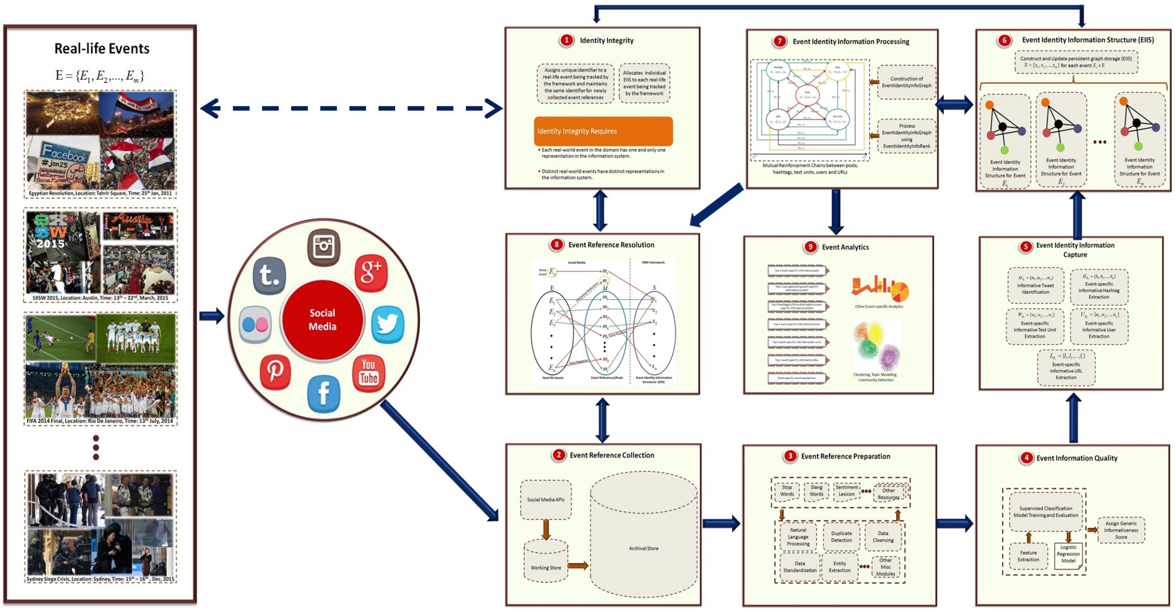
List of Figures

List of Tables

Dedicated to my parents, wife and my entire family for their endless love, support and encouragement.

Dissertation Overview

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



Real-life Events
E = {E1, E2,..., Em}
Egyptian Revolution, Location: Tahrir Square, Time: 25th Jan, 2011
SXSW 2015, Location: Austin, Time: 13th – 22nd March, 2015
FIFA 2014 Final, Location: Rio De Janeiro, Time: 13th July, 2014
Sydney Siege Crisis, Location: Sydney, Time: 15th – 16th, Dec, 2015
Social Media
1 Identity Integrity
Assigns unique identifier to a real-life event being tracked by the framework and maintains the same identifier for newly collected event references
Allocates individual EIIS to each real-life event being tracked by the framework
Identity Integrity Requires
• Each real-world event in the domain has one and only one representation in the information system.
• Distinct real-world events have distinct representations in the information system.
7 Event Identity Information Processing
Construction of EventIdentityInfoGraph
Process EventIdentityInfoGraph using EventIdentityInfoRank
Mutual Reinforcement Chains between posts, hashtags, text units, users and URLs
6 Event Identity Information Structure (EIIS)
Construct and Update persistent graph storage (EIIS) S = {s1, s2,..., sm} for each event Ei ∈ E
Event Identity Information Structure for Event E1
Event Identity Information Structure for Event E2
Event Identity Information Structure for Event Em
8 Event Reference Resolution
Social Media
EIIM Framework
Real-life Events
Event References/Posts
Event Identity Information Structures (EIIS)
9 Event Analytics
Other Event-specific Analytics
Clustering, Topic Modeling, Community Detection
5 Event Identity Information Capture
Informative Tweet Identification
Event-specific Informative Hashtag Extraction
Event-specific Informative Text Unit Extraction
Event-specific Informative User Extraction
Event-specific Informative URL Extraction
2 Event Reference Collection
Social Media APIs
Working Store
Archival Store
3 Event Reference Preparation
Stop Words
Slang Words
Sentiment Lexicon
Other Resources
Natural Language Processing
Duplicate Detection
Data Cleansing
Data Standardization
Entity Extraction
Other Misc Modules
4 Event Information Quality
Supervised Classification Model Training and Evaluation
Feature Extraction
Logistic Regression Model
Assign Generic Informativeness Score

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 Publications

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

- **Debanjan Mahata** and John R. Talburt; *A Framework for Collecting and Managing Entity Identity Information from Social Media.* 19th 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)

Chapter 1

Introduction

1.1 Social Media and Real-life Events

1.2 General Challenges in Mining Social Media Text

1.2.1 Information Overload

A daily average of 58 million tweets is posted in Twitter[1].On an average 60 million photos are shared in Instagram daily[2]. Facebook stores 300 petabytes of data related to its users from all over the world[3]. These are some compelling statistics that makes social media not only rich in volume of data, but also variety, and the velocity at which data is being generated. Due to the great pace at which data is produced in social media, the search engines and content filtering algorithms often face the problem of information overload [1]. They suffer from the dilemma of assessing the accuracy and quality of information content in the sources being produced over their freshness. Thus, collecting different types of references of entities from various social media platforms, assessing their quality, resolving and extracting identity information of the entities poses great challenges in such a situation.

1.2.2 Veracity of Sources

Judging the accuracy of the information and deciding relevant information content in social media references for the purpose of extracting entity identity attributes constitutes

[1] http://www.statisticbrain.com/twitter-statistics/

[2] http://instagram.com/press/

[3] http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/

another challenging situation. For trending topics the search engines have started showing real-time feeds from social media websites in their search results. This has attracted spammers who post trending hash-tags or keywords along with their spam content in order to attract people to their websites offering products or services [2]. An alarming 355% growth of social spam has been reported in 2013[4]. Social media has also been instrumental in spreading misinformation and rumors. Spread of misinformation not only results in pandemonium among the users[5] but also result in extraction of completely wrong information about entities.

1.2.3 Informal Text

Unlike sources of news media and edited documents on the web, the textual content of the social media sources are highly colloquial and pose great difficulties in extracting information. One of the most important sources of information about events, prevalent in the domain of social media are the micro-blogging platforms. Micro blogs pose additional challenges due to their brevity, noisiness, idiosyncratic language, unusual structure and ambiguous representation of discourse [3]. Variation in language, less grammatical structure of sentences, unconventional uses of capitalization, frequent use of emoticons, and abbreviations have to be dealt by any system processing social media content. Moreover, various signals of communications embedded in the text in the form of hash-tags (eg.#sochi), retweets (RT) and user mentions (@) should be understood by the system in order to extract the contextual information hidden in the text. Intentional misspellings sometimes demonstrate examples of intonation in written text [4]. For instance, expressions like, ‘this is so cooool’, emphasizes stress on the emotions and conveys more information that should be captured. It has been shown that it is extremely challenging for the state-of-the art information extraction algorithms to perform efficiently and give accurate results for micro-blogs [5]. For example, named entity recognition methods typically show 85-90% accuracy on longer texts, but 30-50% on tweets [6]. Status messages in social networking websites, content in question answering websites, reviews, and discussions in blogs, and forums exhibit similar nature and present similar challenges to information extraction and text mining procedures.

1.2.4 Sampling Bias

Most commonly used method for obtaining data samples from social media websites is by using their application programming interfaces (APIs). Given the humungous amounts

[4] http://www.likeable.com/blog/2013/11/10-surprising-social-media-statistics/

[5] http://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter

of data produced in real-time, the APIs cannot provide all the data to every single API requests. The requests are often made through a query interface by passing certain query parameters to the APIs. The amount of data returned for against the queries may vary. This depends upon the popularity of the content related to the query. For example, in Twitter studies have estimated that by using Twitter's Streaming API users can expect to receive anywhere from 1% of the tweets to over 40% of tweets in near real-time[6]. The only way to get access to all the tweets is to buy the firehose service, which is seldom done for academic purposes. Other real-time social media publishing services mostly follow the same model. Therefore, this might lead to biasness in the samples collected for studying event related phenomenon and for tracking all the important event related information being produced in real-time.

1.2.5 Multiple Data Sources

The APIs (Application Programming Interfaces) of the different social media websites returns data in different formats (JSON, XML) using different web standards (REST, HTTPS). Moreover, the information obtained from a social media website is dependent upon the type of content it produces. A video sharing website might return an entirely different set of information from a blogging website. Thus, integrating the data obtained from the various social media platforms for the purpose of extraction and tracking of event related information is also one of the challenges.

1.2.6 Lack of Evaluation Datasets

There is a lack of ground truth evaluation data for most of the social media text mining tasks. In traditional data mining research, there is often two types of datasets. One of them is known as training dataset and the other is known as test dataset. The models are trained or developed using the training datasets and are evaluated on test datasets. Thus, the test datasets act as the ground truth. The test dataset for various text mining tasks is mostly not available for social media data. It is often the duty of the researchers to create new test datasets in order to solve a specific task in social media. Sometimes this data might not be a benchmark dataset due to various unwanted noise and human error or perception in annotating the data. This might lead to wrong assumptions and false results.

[6] https://www.brightplanet.com/2013/06/twitter-firehose-vs-twitter-api-whats-the-difference-and-why-should-you-care/

1.3 Research Questions

1.4 Research Methodology

1.5 Research Contributions

1.6 Structure of the Thesis

Chapter 2

Literature Review

2.1 Event Identification in News Text

2.2 Event Identification in Social Media

2.3 Information Quality in Social Media

2.4 Ranking and Summarization of Short Textual Social Media Posts

2.5 Reference Tracking and Entity Resolution

Chapter 3

Defining Events

3.1 Topic Detection and Tracking

3.2 Automatic Content Extraction

3.3 Multimedia Event Detection

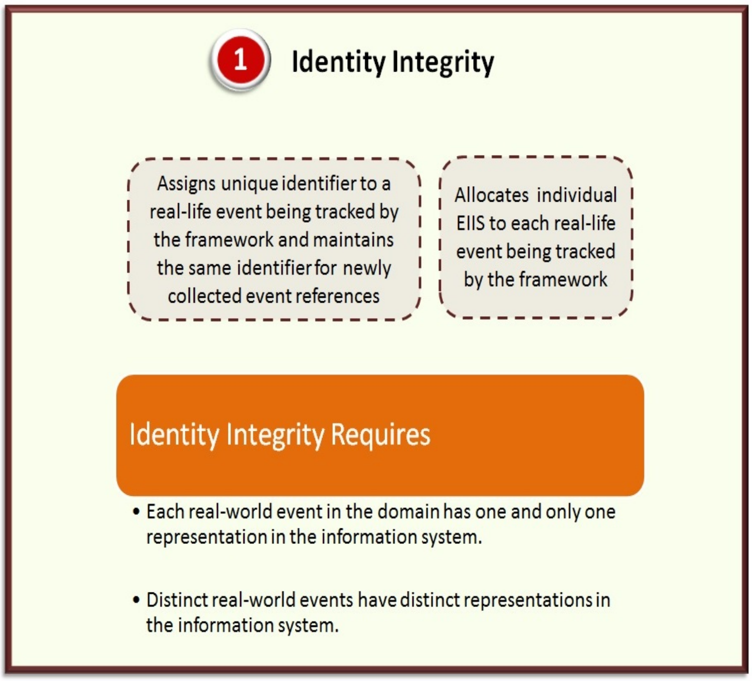
3.4 Events in Social Media

Chapter 4

Event Identity Information Management Life Cycle

4.1 Identity Integrity

Figure 4.1: Identity Integrity component of the EIIM life cycle.



1 Identity Integrity

Assigns unique identifier to a real-life event being tracked by the framework and maintains the same identifier for newly collected event references

Allocates individual EIIS to each real-life event being tracked by the framework

Identity Integrity Requires

- Each real-world event in the domain has one and only one representation in the information system.
- Distinct real-world events have distinct representations in the information system.

FIGURE 4.2: Event Reference Collection component of the EIIM life cycle.

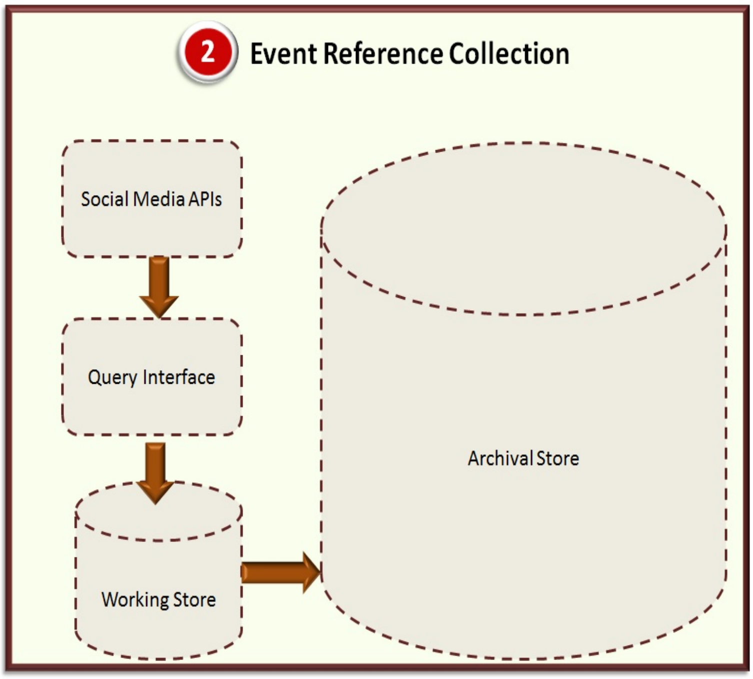
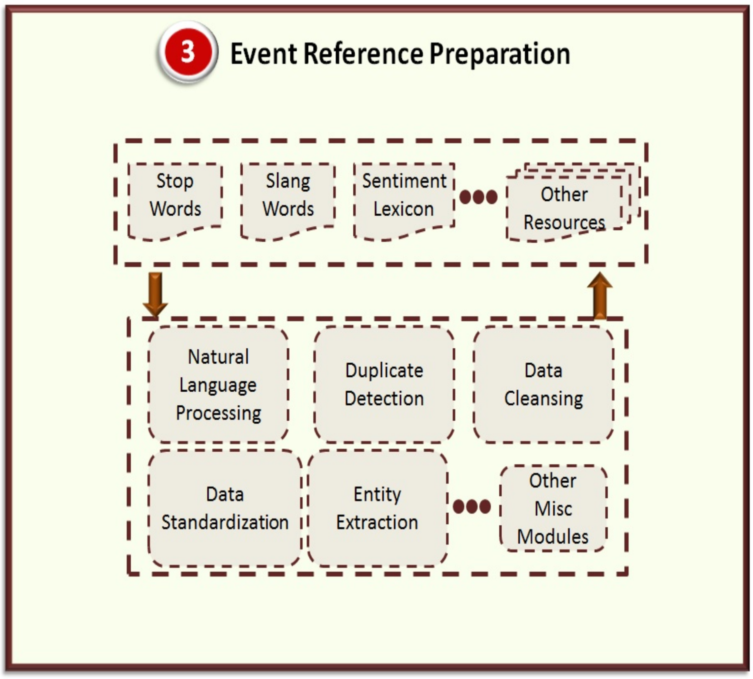


FIGURE 4.3: Event Reference Preparation component of the EIIM life cycle.

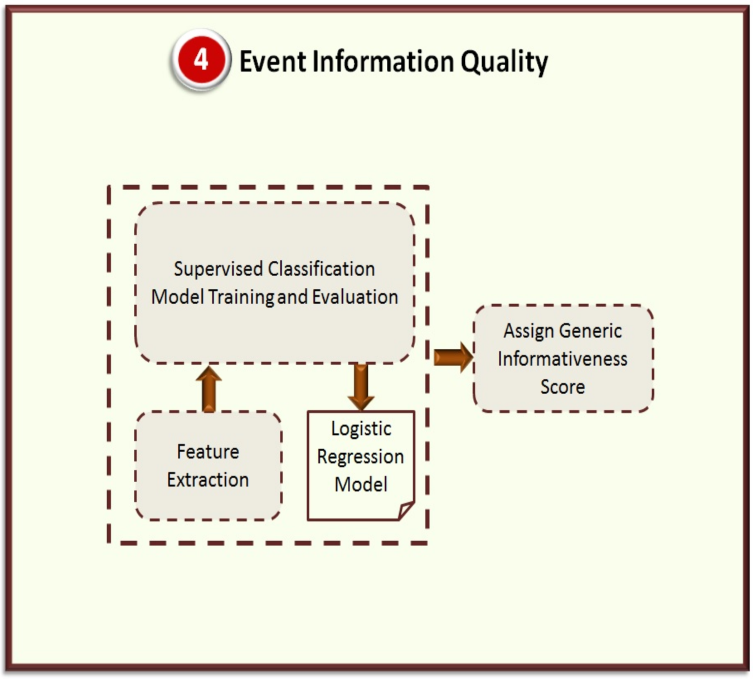


4.2 Event Reference Collection

4.3 Event Reference Preparation

4.4 Event Information Quality

Figure 4.4: Event Information Quality component of the EIIM life cycle.


4
Event Information Quality
Supervised Classification Model Training and Evaluation
Feature Extraction
Logistic Regression Model
Assign Generic Informativeness Score

4.5 Event Identity Information Capture

4.6 Event Identity Information Structure

4.7 Event Identity Information Processing

4.8 Event Reference Resolution

4.9 Event Analytics

Figure 4.5: Event Identity Information Capture component of the EIIM life cycle.

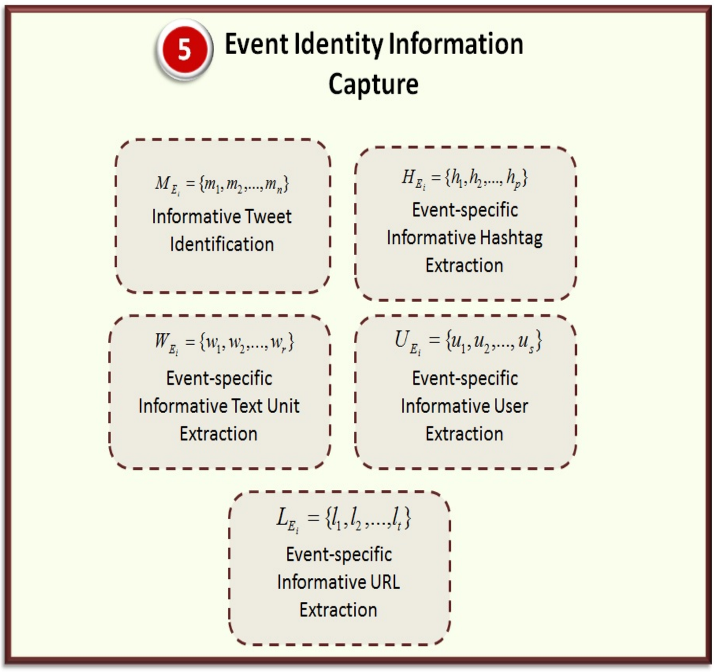

5
Event Identity Information Capture
$M_{E_i} = \{m_1, m_2, ..., m_n\}$
Informative Tweet Identification
$H_{E_i} = \{h_1, h_2, ..., h_p\}$
Event-specific Informative Hashtag Extraction
$W_{E_i} = \{w_1, w_2, ..., w_r\}$
Event-specific Informative Text Unit Extraction
$U_{E_i} = \{u_1, u_2, ..., u_s\}$
Event-specific Informative User Extraction
$L_{E_i} = \{l_1, l_2, ..., l_t\}$
Event-specific Informative URL Extraction

Figure 4.6: Event Identity Information Structure component of the EIIM life cycle.

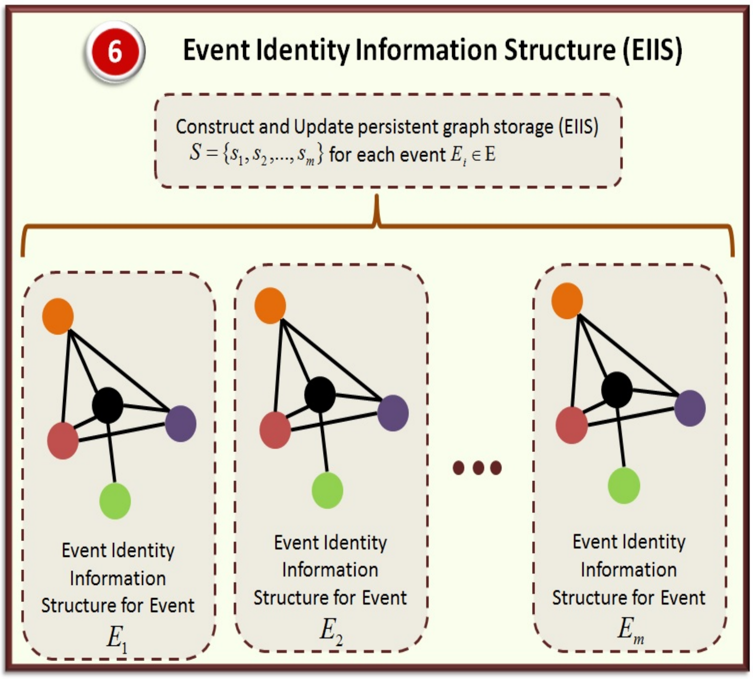
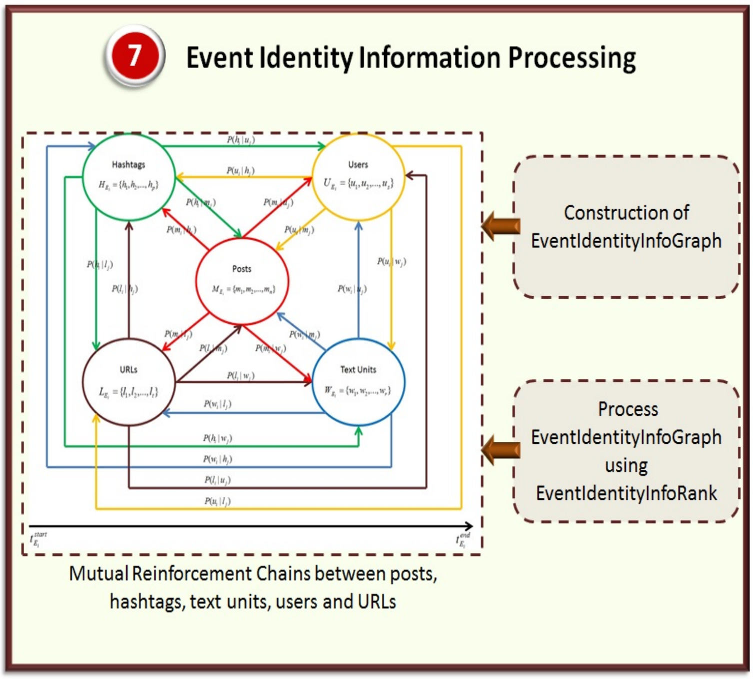

6
Event Identity Information Structure (EIIS)
Construct and Update persistent graph storage (EIIS)
$S = \{s_1, s_2, ..., s_m\}$ for each event $E_i \in \mathrm{E}$
Event Identity Information Structure for Event E_1
Event Identity Information Structure for Event E_2
Event Identity Information Structure for Event E_m

Figure 4.7: Event Identity Information Processing component of the EIIM life cycle.



7
Event Identity Information Processing
Hashtags
$H_{E_i} = \{h_1, h_2, ..., h_p\}$
Users
$U_{E_i} = \{u_1, u_2, ..., u_s\}$
Posts
$M_{E_i} = \{m_1, m_2, ..., m_n\}$
URLs
$L_{E_i} = \{l_1, l_2, ..., l_t\}$
Text Units
$W_{E_i} = \{w_1, w_2, ..., w_r\}$
$P(h_i | u_j)$
$P(u_i | h_j)$
$P(h_i | m_j)$
$P(m_i | u_j)$
$P(m_i | h_j)$
$P(u_i | m_j)$
$P(h_i | l_j)$
$P(l_i | h_j)$
$P(u_i | w_j)$
$P(w_i | u_j)$
$P(m_i | l_j)$
$P(l_i | m_j)$
$P(m_i | w_j)$
$P(w_i | m_j)$
$P(l_i | w_j)$
$P(w_i | l_j)$
$P(h_i | w_j)$
$P(w_i | h_j)$
$P(l_i | u_j)$
$P(u_i | l_j)$
$t_{E_i}^{start}$
$t_{E_i}^{end}$
Construction of EventIdentityInfoGraph
Process EventIdentityInfoGraph using EventIdentityInfoRank
Mutual Reinforcement Chains between posts, hashtags, text units, users and URLs

Figure 4.8: Event Reference Resolution component of the EIIM life cycle.

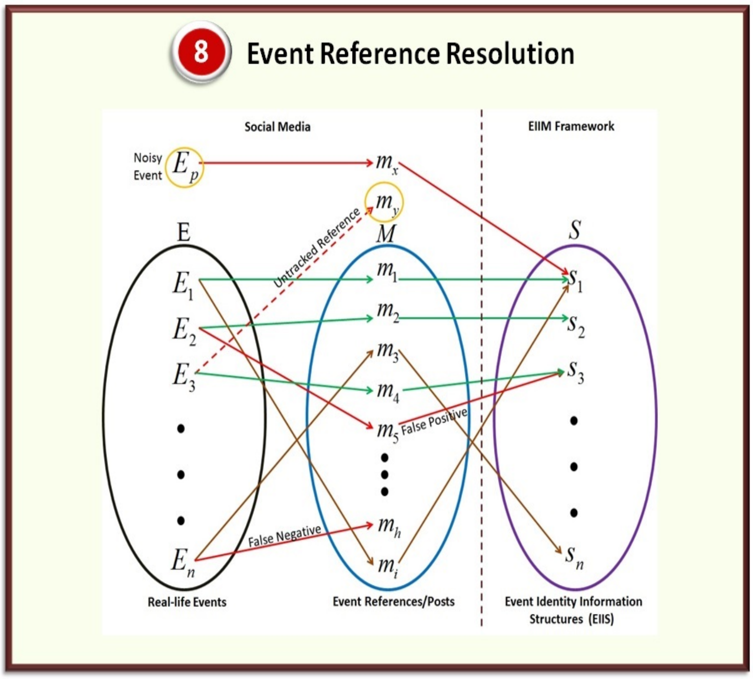
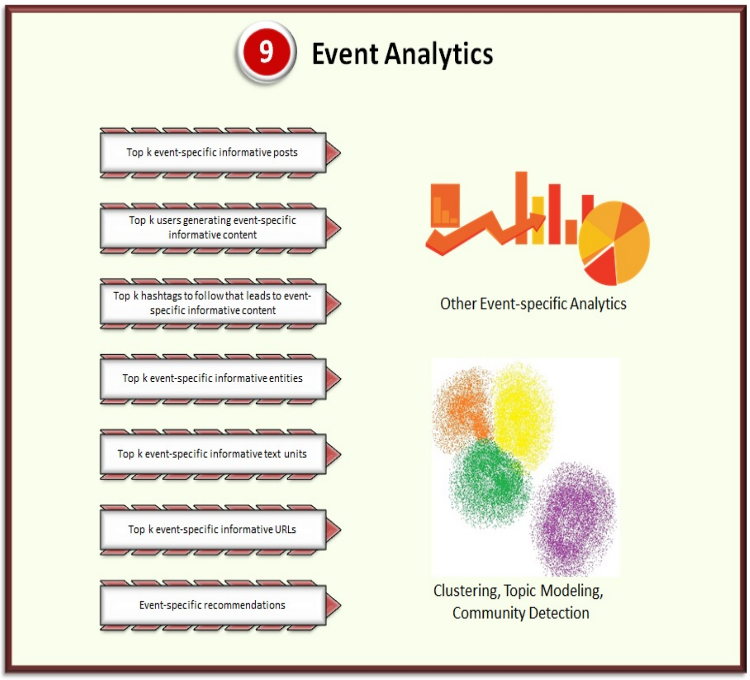


Figure 4.9: Event Analytics component of the EIIM life cycle.


9
Event Analytics
Top k event-specific informative posts
Top k users generating event-specific informative content
Top k hashtags to follow that leads to event-specific informative content
Top k event-specific informative entities
Top k event-specific informative text units
Top k event-specific informative URLs
Event-specific recommendations
Other Event-specific Analytics
Clustering, Topic Modeling, Community Detection

Chapter 5

Discovering Event-specific Informative Content from Twitter

Twitter has brought a paradigm shift in the way we produce and curate information about real-life events. Huge volumes of user-generated tweets are produced in Twitter, related to events. Not, all of them are useful and informative. A sizable amount of tweets are spams and colloquial personal status updates, which does not provide any useful information about an event. Thus, it is necessary to identify, rank and segregate event-specific informative content from the tweet streams. In this chapter, we implement *EventIdentityInfoGraph* and *EventIdentityInfoRank* as introduced in 4.7 in the context of Twitter. We name *EventIdentityInfoGraph* as *TwitterEventInfoGraph* and *EventIdentityInfoRank* as *TwitterEventInfoRank*. Mutually reinforcing relationships between tweets, hashtags, text units, URLs and users are defined and represented using *TwitterEventInfoGraph*. *TwitterEventInfoRank* simultaneously ranks tweets, hashtags, text units, URLs and users producing them in terms of event-specific informativeness by leveraging the semantics of relationships between each of them as represented by *TwitterEventInfoGraph*. Experiments and observations are reported on four million (approx) tweets collected for five real-life events, and evaluated against popular baseline techniques showing significant improvement in performance.

5.1 Twitter and Event Related Content

Social media platforms provide multiple venues to people for sharing first-hand experiences and exchange information about real-life events. Twitter is one such platform that has become an indispensable source for disseminating news and real-time information about current events. It is a microblogging application that allows its users to post short messages of 140 characters known as tweets, from a variety of internet enabled devices. Studies have shown the importance of Twitter as a news circulation service [7], and a source for gauging public interest and opinions [8]. 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 [9–11].

Table 5.1: Examples of different event related tweets.

Ted Cruz is a dangerous man. Crazy and gaining support. Megalomaniac leaders are bad, mkay. #CPAC #politics #joke [***personal/uninformative***] ***Event: ‘CPAC 2014’***
Thanks for the memories Sochi! I've had the time of my life #Sochi2014 #sochiselfie http://t.co/DqkLEaAMpo. [***personal/uninformative***] ***Event: ‘Sochi Games’***
#SXSW14 #SXSW #sxswinteractive #CPAC2014 #CPAC #CPACPickupLines #CPACPanels Be squared away perky TOP TWEETED of http://t.co/h0igdOVNW0. [***spam/uninformative***] ***Event: ‘CPAC 2014’***
In #Sochi, the Dutch are dominating the overall Olympic medal count http://t.co/jMR1WUqEK4 (Reuters) http://t.co/dAfDhEgTGA. [***event-specific informative***] ***Event: ‘Sochi Games’***
New post: Sochi Was For Suckers - Laugh Studios/ http://t.co/cWQJCBp3Ow #lol #funny #rofl #funnypic #fail #wtf. [***spam/uninformative***] ***Event: ‘Sochi Games’***
It's tedcruz vs. SenJohnMcCain in a #CPAC spat. What did they say? Find out on #AC360 8p on CNN. [***event-specific informative***] ***Event: ‘CPAC 2014’***

Users not only post plain textual content in their messages but also share URLs, linking to other external websites, images and videos. Apart from creating new content, the users also share content produced by others. This activity is known as *retweeting*, and such tweets are preceded by special characters ‘*RT*’. The messages are normally written by a single person and are read by many. The readers in this context are known as *followers*, and the user whom they follow is considered as their *friend*. Any user with good intent either share messages that might be of interest to his followers, or for joining conversations on topics of his interest. The ‘@’ symbol followed by the username commonly known as *user mentions*, is used for mentioning other users in tweets for initiating conversations.

The concise and informal content of a tweet is often contextualized by the use of a crowdsourced annotation scheme called *hashtags*. Hashtags are a sequence of characters in any language prefixed by the symbol ‘#’ (for e.g. #websci2015). They are widely used by the users for categorizing the content based on a topic, join conversations related to a topic, and to make the tweets easily searchable by other interested users. They also act as strong identifiers of topics [12]. When tweeting about real-life events the users also tend to use hashtags in order to post event-specific content. For e.g. ‘#Egypt’ and ‘#Jan25’, were among the most popular hashtags in Twitter used for spreading, organizing and analyzing information related to ‘Egyptian Revolution of 2011’ [13].

284 million monthly users of Twitter posting 500 million tweets per day produces a variety of content[1]. A significant proportion of it are related to different real-life events (e.g, football matches, conferences, music shows, etc). Majority of this content are personal updates (e.g. *Thanks for the memories Sochi! I've had the time of my life #Sochi2014 #sochiselfie http://t.co/DqkLEaAMpo*), pointless babbles (e.g. *Ted Cruz is a dangerous man. Crazy and gaining support. Megalomaniac leaders are bad, mkay. #CPAC #politics #joke*) and spams (e.g *New post: Sochi Was For Suckers - Laugh Studios/ http://t.co/cWQJCBp3Ow #lol #funny #rofl #funnypic #wtf.*). Personal views and conversations might be of interest to a specific group of people. However, they are meaningless and provides no information to the general audience. On the other hand there are tweets that presents newsworthy content, recent updates and real-time coverage of on-going events (e.g. *In #Sochi, the Dutch are dominating the overall Olympic medal count http://t.co/jMR1WUqEK4 (Reuters) http://t.co/dAfDhEgTGA*). These tweets provide event-specific informative content and are more useful for general audience interested to know about the event. We call them as event-specific informative tweets. Table 5.1 presents some examples of different types of tweets shared during real-life events.

5.2 Motivation

With the plethora of event related content being produced in Twitter, it becomes inconvenient for users to search and follow informative posts. This necessitates development of techniques that can identify and rank tweets in terms of their event-specific informativeness. In addition to the tweets, a backend automated system dedicated for processing, analyzing and presenting information from Twitter during an event, could get immensely benefitted from identification and ranking of event-specific informative hashtags, text units, users and URLs. This would enable the system to generate answers to questions like: *Who are the users producing large amount of event-specific informative content?. Which are the best hashtags and URLs to follow that would lead to high quality event-specific information?. Which are the best hashtags and text units to index for efficient retrieval of event-specific information?.* Such a system would further facilitate better consumption of content while exploring event information from Twitter. It could have a positive impact on triggering event-specific recommendations and efficient processing of information. It can act as a core component of event management, event summarization, event marketing and journalistic platforms leveraging Twitter.

[1] http://about.twitter.com/company

5.3 Challenges in Mining Tweets

Apart from the problem of information overload, microblogging websites like Twitter pose challenges for automated information mining tools and techniques due to their brevity, noisiness, idiosyncratic language, unusual structure and ambiguous representation of discourse. Information extraction tasks using state-of-the-art natural language processing techniques, often give poor results for tweets [6]. Abundance of link farms, unwanted promotional posts, and nepotistic relationships between content creates additional challenges. Due to the lack of explicit links between content shared in Twitter it is also difficult to implement and get useful results from ranking algorithms popularly used for web pages. Lastly, to our knowledge, there is an absence of techniques at present that is capable of simultaneously ranking and identifying event-specific informative tweets, hashtags, text units, users and URLs, with an ability to scale.

5.4 Objective and Contributions

The main objective of our work is to automatically identify and rank event-specific informative content posted in Twitter. Our primary hypothesis is that there are explicit cues available in the content of the tweets posted during an event for determining event-specific informativeness. Our approach is based on the *principle of mutual reinforcement* commonly used for summarization of textual documents. We build our methodology on the basic tenets of *Mutually Reinforcing Chains* [14], for ranking and identification of event-specific informative content in Twitter. We make the following contributions:

- analysis of informative and non-informative content in 3.8 million event related tweets;
- propose a generic framework based on principle of mutual reinforcement that takes into account the semantics of relationships between *tweets*, *hashtags*, *text units*, *URLs* and *users*, and represent them in a graph structure - *TwitterEventInfoGraph*;
- leverage the mutually reinforcing relationships in *TwitterEventInfoGraph* and develop a graph based iterative algorithm - *TwitterEventInfoRank*, for simultaneously ranking *tweets*, *hashtags*, *text units*, *users* and *URLs* in terms of event-specific informativeness;
- evaluate the algorithm against popular baselines and report its performance in identifying and ranking event-specific informative content from Twitter.

5.5 Analysis of Informative and Non-informative Content in Tweets

5.6 EventIdentityInfoGraph

5.7 EventIdentityInfoRank

5.8 Experiments

5.8.1 Data Collection

5.8.2 Data Preparation

5.8.3 Baselines

5.8.4 Evaluation

5.8.5 Sample Results

Chapter 6

Potential Applications of the EIIM Framework

6.1 Event Monitoring and Analysis

References related to real-life events are extremely abundant in social media. Right from natural disasters such as the 'Haiti Earthquake' [15] to international sporting events like the 'Winter Olympics' [16] to socio-political [17] and socio-economical [18] events that shook the world such as presidential elections [19], 'Egyptian Revolution' [20], and recessions were covered, analyzed, extrapolated and informed by social media. This prolific event-specific content in social media makes it a promising ground for performing event analytics. Platforms like Geofeedia[1], TwitterStand[2], Twitris[3], Truthy[4], and Tweet-Tracker[5] have developed techniques to provide analytics related to different local and global real-life events.

Monitoring social media has become one of the essential activities of national security agencies for predicting potential threats and mass protests [21]. Social media is being used for tracking terrorism activities [22], collective actions [23], and countering cyber-attack threats[6]. One of the main components of each of these applications is tracking references related to the events. The proposed EIIM model could be an essential component of such systems. It would help in identifying, tracking and analyzing events and its related references in an organized manner over time.

[1] http://geofeedia.com/

[2] http://twitterstand.umiacs.umd.edu/

[3] http://twitris.knoesis.org/

[4] http://truthy.indiana.edu/

[5] http://tweettracker.fulton.asu.edu/

[6] https://www.recordedfuture.com/

6.2 Event Information Retrieval

Retrieving informative content related to real-life events shared in social media and presenting them in an organized way to the interested users has led to web based services like Seen[7]. It allows users to follow live updates of the events and also aids in witnessing and re-living the events at a later stage from the archives. Showing useful and interesting content to users by filtering out the pointless babbles from social media streams is an important component of such services. Additionally, such systems could get immensely benifitted by identification of event-specific informative hashtags, text units, users and URLs over time as the event proceeds. This would further enable efficient indexing of event-specific terms and hashtags that leads to high quality information, and effective processing of information. It would enhance the user experience, allowing better consumption and summarization of information related to the events, and positively impact triggering of event-specific recommendations. Thus, the proposed EIIM model in this thesis can act as the core component of information retrieval systems retrieving and organizing information related to real-life events from social media.

6.3 Opinion and Review Mining

Every day millions of people express their opinions in social media about products and companies they like and dislike. Their communications often include thoughts about good and bad experiences with the products and services. This provides a great opportunity for companies to understand its customers and to get unbiased valuable feedback from them about their product offerings without asking them to fill out time consuming outdated surveys. The EIIM framework when used for monitoring references of products/services from social media during product launch events could be useful in mining isightful and informative opinionated content. Combined with sentiment analysis, the invention could be a powerful tool for review analysis. One of the important contributions of the system could be to identify the sources having high chances of containing insightful information and filter them out for further processing. This would make a review mining system more efficient and increase its overall quality. Mining opinions related to entities related to an event could be used in many other contexts like political campaigns, socio-political studies, market behavior analysis, e-commerce applications, etc. Steps are being taken for adding this capability to the EIIM framework. On considering a mix of named entities and unigram opinionated words as text units in the *EventIdentityInfoGraph* we obtained some preliminary encouraging results. A glimpse

[7] http://seen.co

of the results obtained for a basketball game ”Miami Heats VS Cleveland Cavaliers”, played on 25th December, 2014 is as follows:

Top 10 insightful and opinionated tweets for an hour related to the game

1. Good win for the Heat tonight against Cavs and Lebron. Great game for Wade and Deng. Just imagine if Bosh were healthy. #HeatvsCavs
2. Good work Dwayne Wade. Good work Miami Heat. LeBron is embarrassed. It’s all over his face. #NBA #heatvscavs
3. Great game on Christmas Heat Showed up and spoiled Lebron Return to MIA! #Wade County #HeatvsCavs #NBAChristmas
4. Lebron leaves Miami high and dry and they cheer his return. Some even cheering cavs. Embarrassing bandwagon fan base. #heatv. . .
5. I totally understand LBJ move to Cleveland and like it. But if I’m a #Miami fan, I would boo LeBron like crazy today. #heatvscavs #CLEvsMIA
6. Stay classy #Miami. Good game vs. Lebron and; Cavs. #NBA #MIAvsCLE #HeatvsCavs #Heat #HeatNation
7. Loul Deng playing both ends of the floor. He’s playing good D to LBJ #heatvscavs
8. Heat fans ; Cavs fans. Class vs no class. No burning a jersey in Miami #heatvscavs #HeatNation
9. WE FUCKING WON!!!!!! LETS GO HEAT #HEATgame #HeatNation #HeatvsCavs Wade with 31 points 5 assist 5 rebounds! Good shit MIAMI
10. Kevin Love is overrated. Big fish, small pond in MN and injury prone. #HeatvsCavs #NBAXmas

The above tweets point to the reactions of the viewers on the game as well as the players participating in the event.

6.4 Recommender Systems

The EIIM framework can be used for developing event related recommender systems. The ranked list of event identity information can be used for giving useful recommendations. For example following is a refined tweet recommendation for an event obtained from a snapshot of the *EventIdentityInfoGraph* created for the event: "BlackLivesMatter": Protest movement against the killing of Eric Garner.

Original Tweet:

- #BREAKING #NEWS — New York City Mayor Says, #BlackLivesMatter http://t.co/qYvp8L8gDh — #BLACK HCP520

Recommended Tweets:

- New York: What's the plan? Where are the protests happening tonight? #EricGarner #BlackLivesMatter #MichaelBrown #ICantBreathe
- Brooklyn District Attorney to Convene Grand Jury in Case of #AkaiGurley NBC New York http://t.co/mLlYPy39Pa #BlackLivesMatter
- New York Today! #ShutItDown #economicshutdown #BlackLivesMatter #ICantBreathe #EricGarner #nojusticenoprofits http://t.co/F0TrZtx2Y5

Similarly an user can get other recommended users who are talking on the same topic. Hashtags and topics can also be recommended. It can further lead to clustering of similar content and discovery of communities around different topics related to the event. We wish to work on this in the future.

6.5 Event Management and Marketing

Social media is increasingly being used by event management practitioners while organizing conferences, seminars, music festivals, fashion shows, fundraisers and various other types of planned events. Tracking and producing useful and informative content before, during and after the events in social media from the perspective of event management has proved to be extremely beneficial [8]. Right from promoting the events, collecting RSVPs, creating communities around topics, announcing important information, getting real-time unbiased feedbacks, to marketing right content to the users creating buzz about the events, social media plays an important role. It also helps in building long term relationships with the communities of users interested in an event and track their related activities. In such a scenario the EIIM life cycle can constantly track and persistently store salient information related to events right from its inception. The *EventIdentityInfoGraph* can aid in identifying event-specific informative content and users producing

[8] http://oursocialtimes.com/using-social-media-to-make-your-event-a-dazzling-success-infographic/

them, which could further lead to effective targeting of user communities, generating event summaries, mining opinions, broadcasting interesting information, among other things related to an event.

6.6 Social Media Data Integration

Organizations have increasingly started integrating the data available in social media with the enterprise data[9]. Social media data is most powerful when it is combined with daily transactional data and the master data to give a comprehensive view of customers, products and business conditions. Customers often openly talk about the products in social media and build communities around hashtags [**?**] related to different topics. The EIIM framework could go a long way in collecting right information about the entities of concern maintained in the enterprise databases and integrate the collected information with the already existing ones. The entity resolution aspect would further help in managing the data quality issues related to data integration. In such conditions the EIIM model proposed could be used for integrating entity information from two distinct domains of enterprise system and social media in order to gain strategic intelligence related to business of an organization. This would further help an organization in marketing, corporate communications, public relations, customer support, product development, advertising, market research, product recommendations and gaining competitive intelligence.

[9] http://www.altimetergroup.com/research/reports/social-data-intelligence

Chapter 7

Conclusion and Future Work

7.1 Conclusion

7.2 Future Work

7.2.1 Summarizing Event Related Content

7.2.2 Identifying Insightful Opinionated Content Related to Events

7.2.3 Event Topic Modeling

7.2.4 Event-specific Recommendations

7.2.5 Distributed Processing of EventIdentityInfoGraph

7.2.6 Event Ontology for Social Media

Appendix A

Appendix Title Here

Write your Appendix content here.

Bibliography

[1] Paul Hemp. Death by information overload. *Harvard business review*, 87(9):82–89, 2009.

[2] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, page 12, 2010.

[3] Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A Greenwood, Diana Maynard, and Niraj Aswani. Twitie: An open-source information extraction pipeline for microblog text. In *RANLP*, pages 83–90, 2013.

[4] Scott Prevost. An information structural approach to spoken language generation. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 294–301. Association for Computational Linguistics, 1996.

[5] Leon Derczynski, Diana Maynard, Niraj Aswani, and Kalina Bontcheva. Microblog-genre noise and impact on semantic annotation accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 21–30. ACM, 2013.

[6] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[7] Owen Phelan, Kevin McCarthy, and Barry Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*, pages 385–388. ACM, 2009.

[8] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010.

[9] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *Knowledge and Data Engineering, IEEE Transactions on*, 25(4):919–931, 2013.

[10] Ana-Maria Popescu, Marco Pennacchiotti, and Deepa Paranjpe. Extracting events and event descriptions from twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 105–106. ACM, 2011.

[11] Hemant Purohit and Amit P Sheth. Twitris v3: From citizen sensing to analysis, coordination and action. In *ICWSM*, 2013.

[12] David Laniado and Peter Mika. Making sense of twitter. In *The Semantic Web–ISWC 2010*, pages 470–485. Springer, 2010.

[13] Genevieve Barrons. 'suleiman: Mubarak decided to step down# egypt# jan25 oh my god': Examining the use of social media in the 2011 egyptian revolution. *Contemporary Arab Affairs*, 5(1):54–67, 2012.

[14] Furu Wei, Wenjie Li, Qin Lu, and Yanxiang He. Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 283–290. ACM, 2008.

[15] Huiji Gao, Geoffrey Barbier, Rebecca Goolsby, and Daniel Zeng. Harnessing the crowdsourcing power of social media for disaster relief. Technical report, DTIC Document, 2011.

[16] Shaun Walker. Russia to monitor 'all communications' at winter olympics in sochi. *The Guardian, October*, 6, 2013.

[17] Vivek Kumar Singh, Debanjan Mahata, and Rakesh Adhikari. Mining the blogosphere from a socio-political perspective. In *Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on*, pages 365–370. IEEE, 2010.

[18] Johan Bollen, Alberto Pepe, and Huina Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *arXiv preprint arXiv:0911.1583*, 2009.

[19] Emily Metzgar and Albert Maruggi. Social media and the 2008 us presidential election. *Journal of New Communications Research*, 4(1):141–165, 2009.

[20] Alok Choudhary, William Hendrix, Kathy Lee, Diana Palsetia, and Wei-Keng Liao. Social media evolution of the egyptian revolution. *Communications of the ACM*, 55(5):74–80, 2012.

[21] Jeffrey Ghannam. Social media in the arab world: Leading up to the uprisings of 2011. *Center for International Media Assistance*, 3, 2011.

[22] Onook Oh, Manish Agrawal, and H Raghav Rao. Information control and terrorism: Tracking the mumbai terrorist attack through twitter. *Information Systems Frontiers*, 13(1):33–43, 2011.

[23] Nitin Agarwal, Merlyna Lim, and Rolf T Wigand. *Online collective action: Dynamics of the crowd in social media*. Springer, 2014.