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

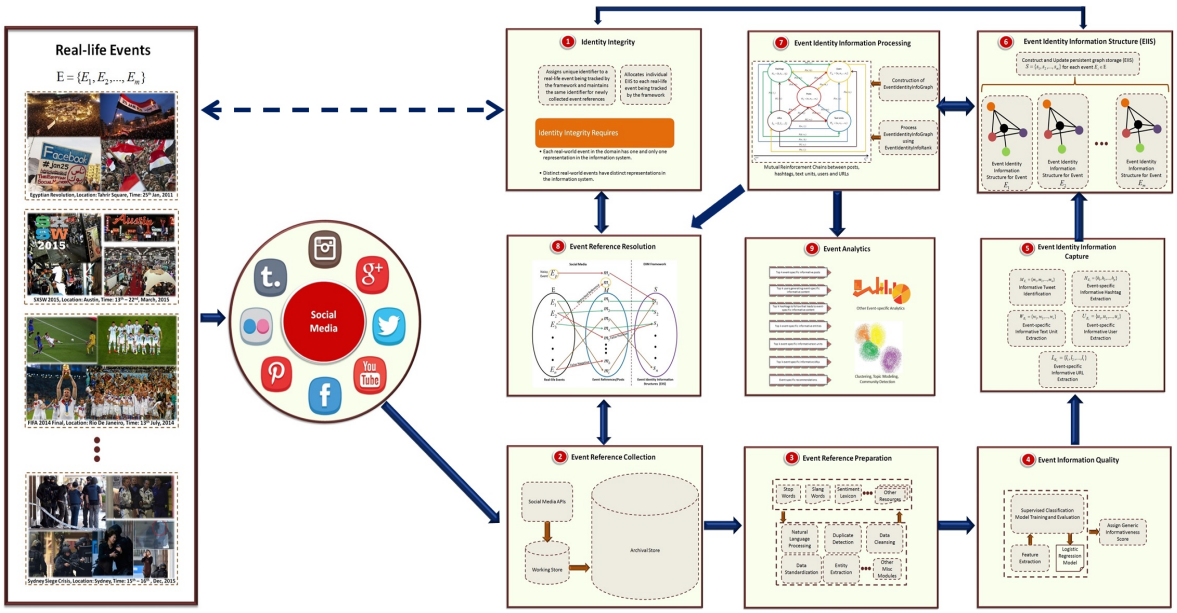
List of Figures

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



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*. 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)

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 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

Defining Events

2.1 Topic Detection and Tracking

2.2 Automatic Content Extraction

2.3 Multimedia Event Detection

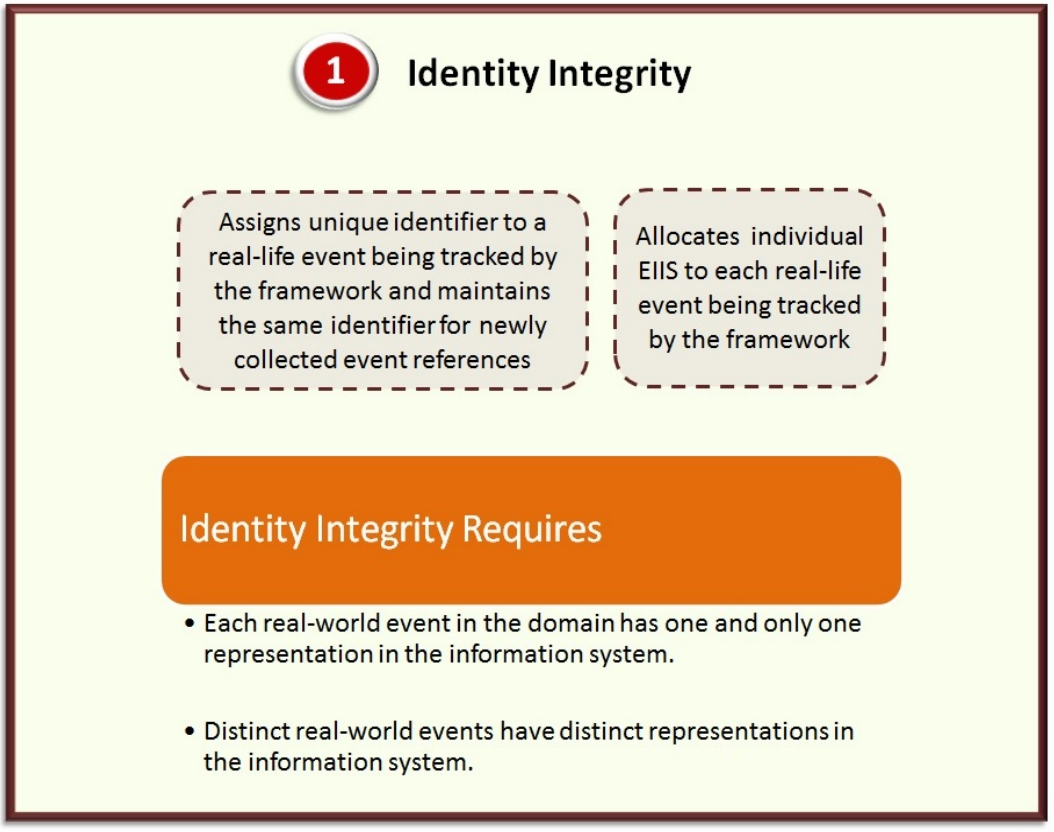
2.4 Events in Social Media

Chapter 3

Event Identity Information Management Life Cycle

3.1 Identity Integrity

FIGURE 3.1: Identity Integrity component of the EIIM life cycle.



3.2 Event Reference Collection

FIGURE 3.2: Event Reference Collection component of the EIIM life cycle.

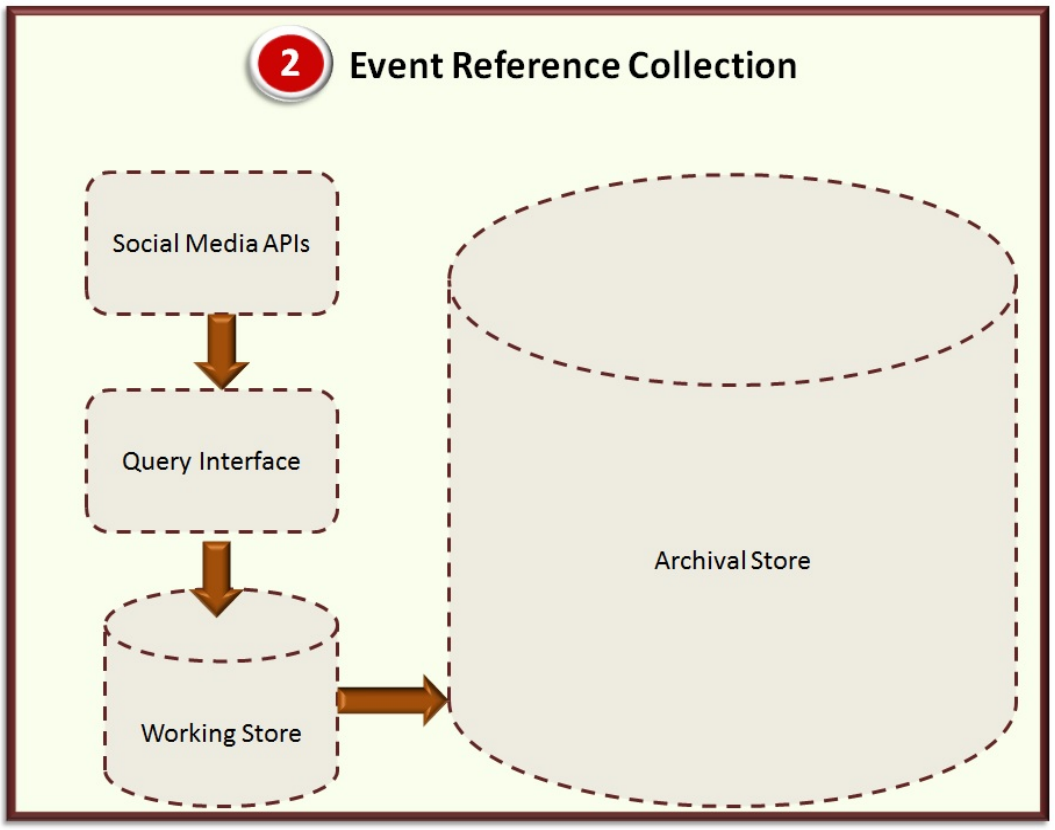


FIGURE 3.3: Event Reference Preparation component of the EIIM life cycle.

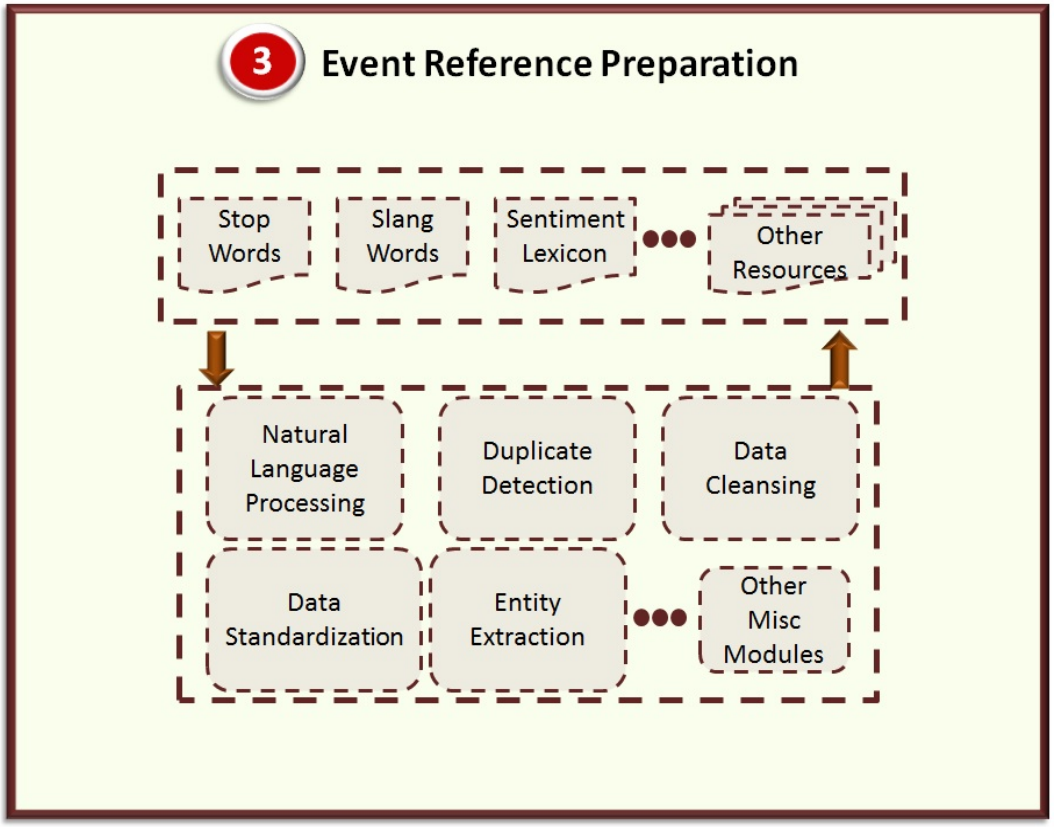


FIGURE 3.4: Event Information Quality component of the EIIM life cycle.

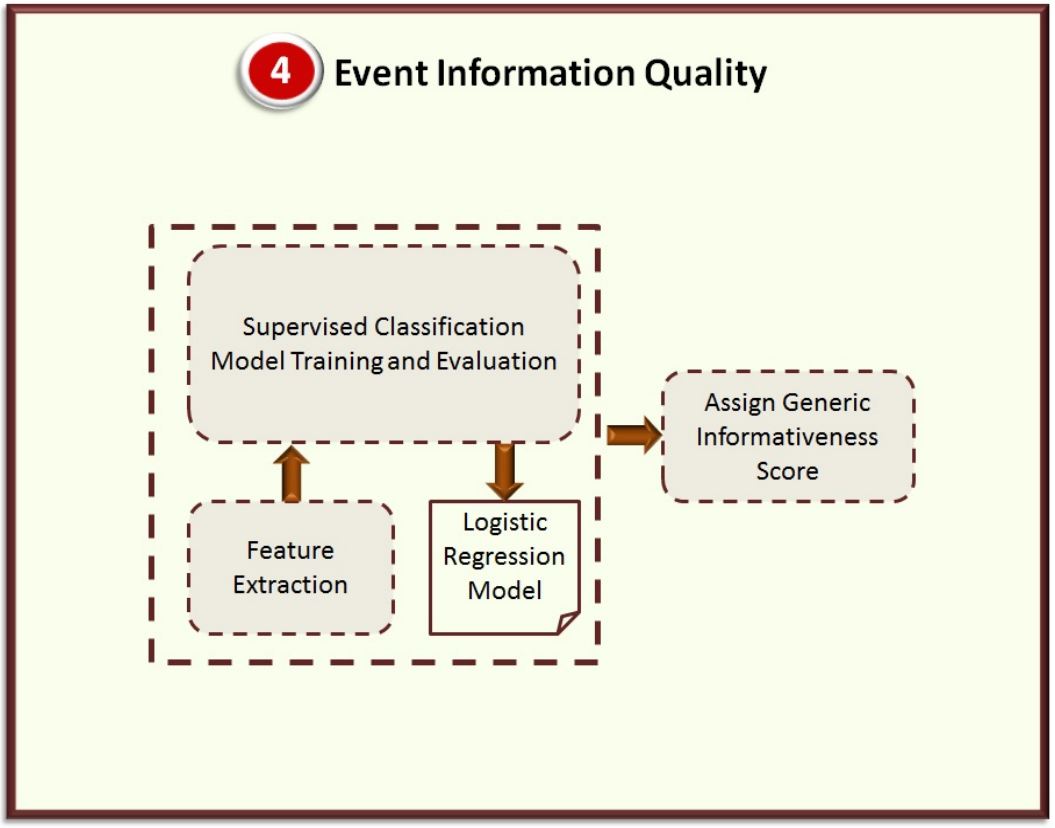


FIGURE 3.5: Event Identity Information Capture component of the EIIM life cycle.

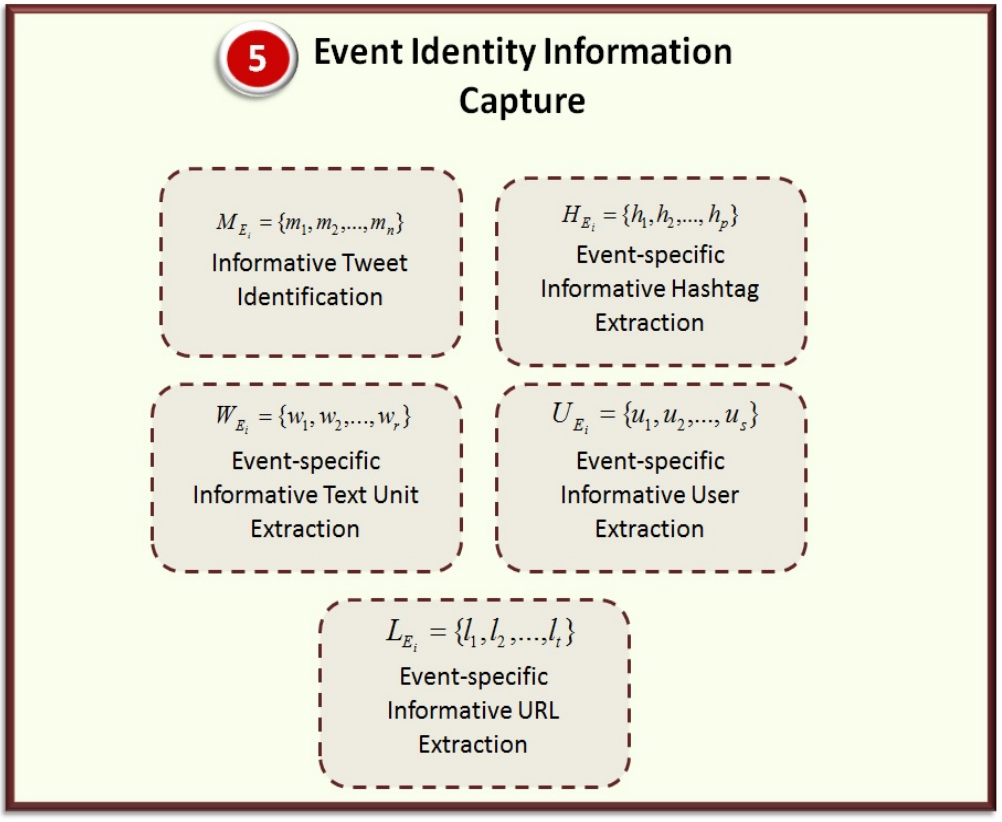


FIGURE 3.6: Event Identity Information Structure component of the EIIM life cycle.

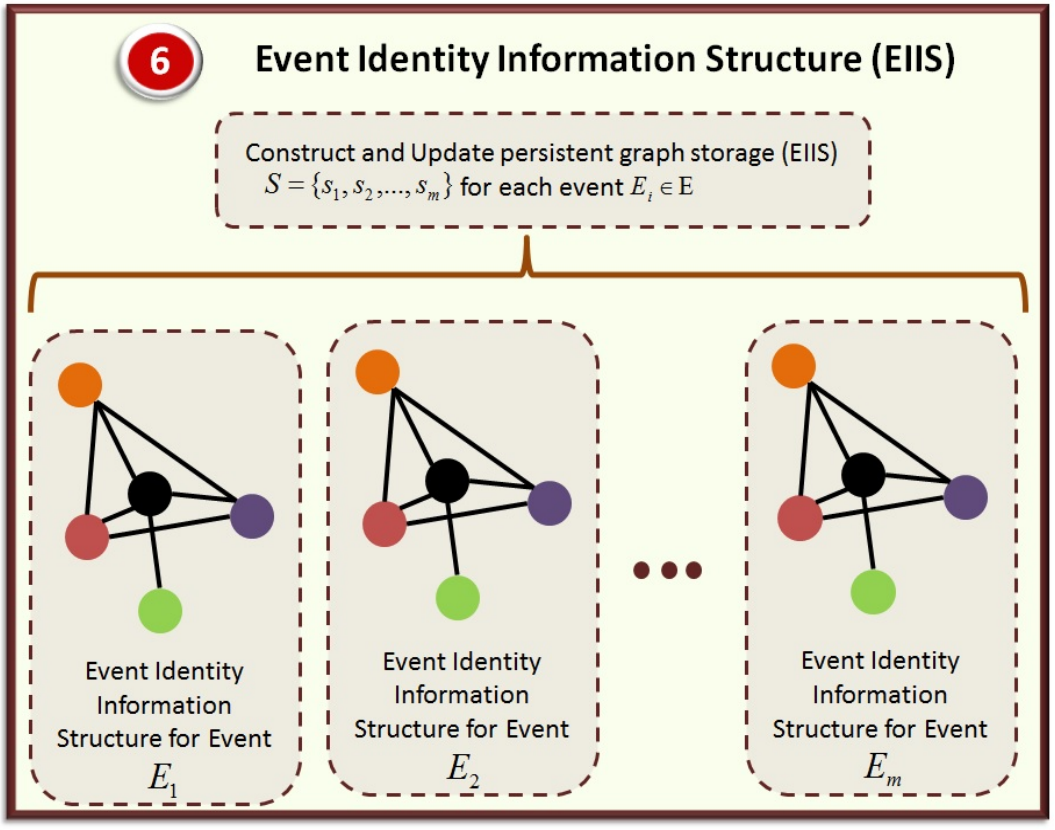


FIGURE 3.7: Event Identity Information Processing component of the EIIM life cycle.

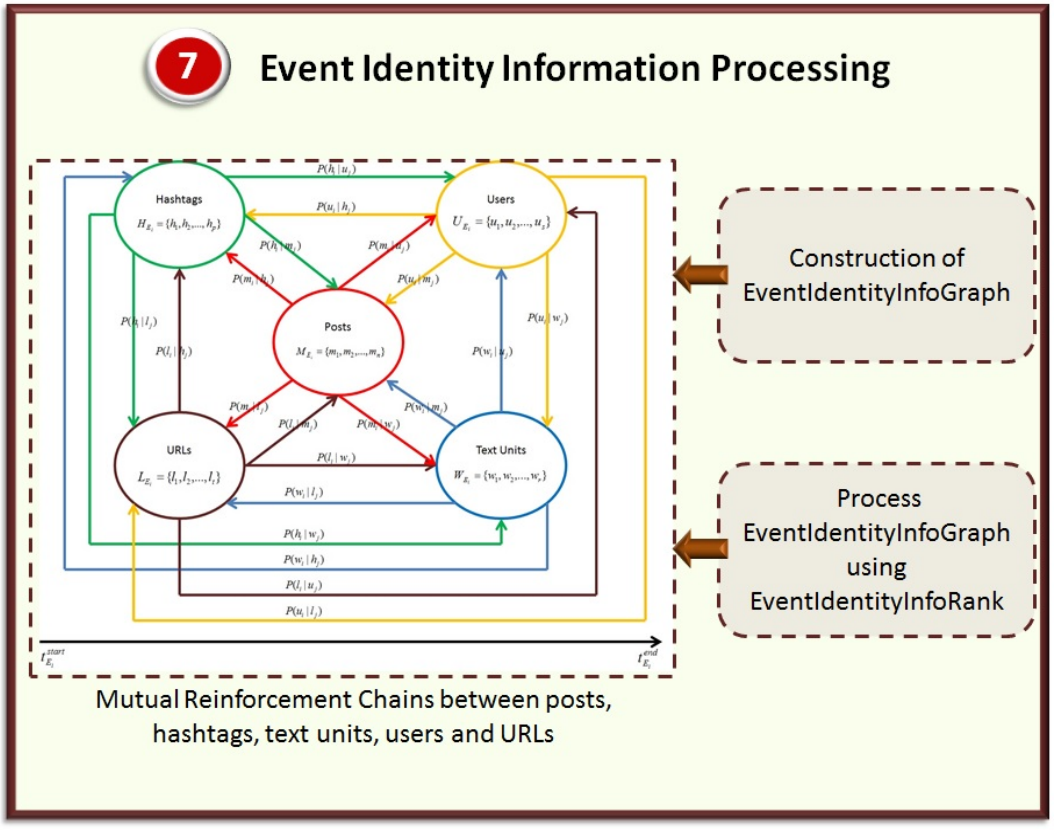
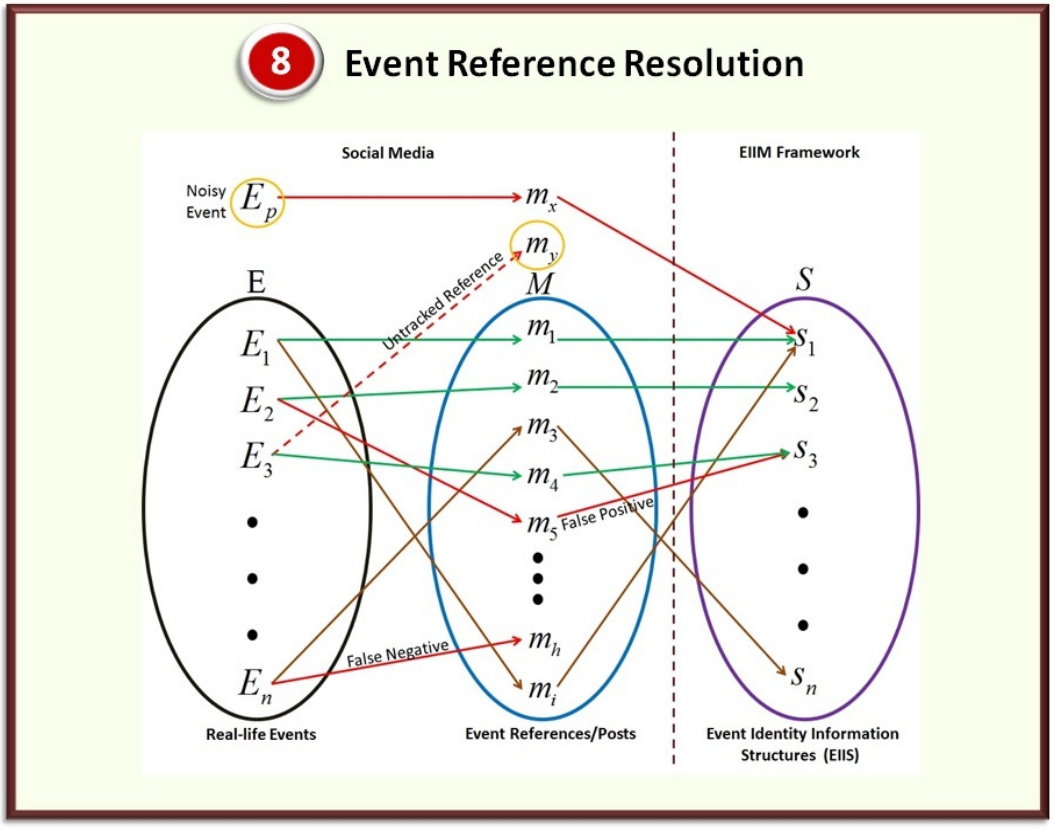


FIGURE 3.8: Event Reference Resolution component of the EIIM life cycle.



3.3 Event Reference Preparation

3.4 Event Information Quality

3.5 Event Identity Information Capture

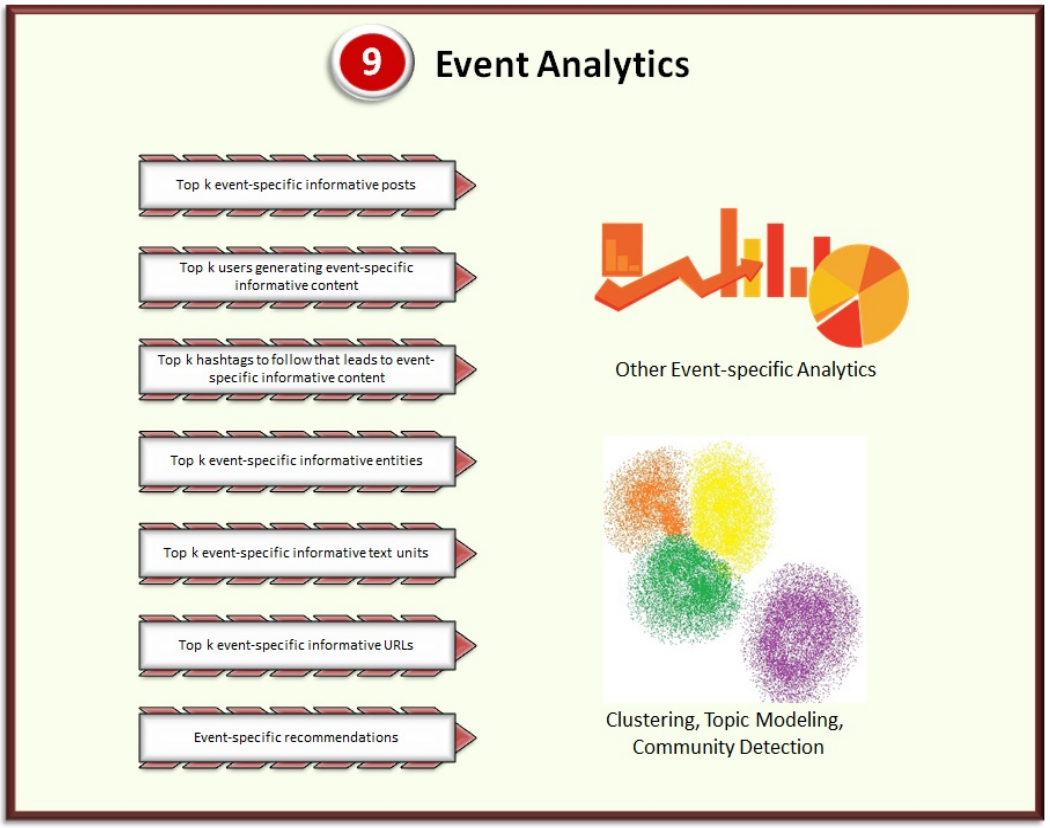
3.6 Event Identity Information Structure

3.7 Event Identity Information Processing

3.8 Event Reference Resolution

3.9 Event Analytics

FIGURE 3.9: Event Analytics component of the EIIM life cycle.



Chapter 4

Identifying Event-specific Sources from the Blogosphere

Abstract

Social media has become a useful medium for mobilizing support for various real-life events and a platform for the public to voice their opinion freely to a huge audience in the web. Social media sources often provide novel and specific information in contrast to the generic information obtained from the mainstream media. This makes social media a valuable source for conducting studies and analyzing events. However, due to the power law distribution of the Internet, these sources get buried in the Long Tail. The overwhelming number of Long Tail social media sources makes it more challenging to identify the valuable sources with specific information. It is, therefore, of utmost importance to identify quality sources from these social media sites for understanding and exploring an event. We propose an evolutionary mutual reinforcement model for identifying and ranking highly 'specific' social media sources, otherwise buried in the Long Tail, and 'close' entities related to an event. We also introduce a novel evaluation strategy, due to the absence of ground truth for validating the results. We observe a huge percentage gain in information between 25% and 130% against the baselines (viz., Google search and Icerocket blog search) from the sources ranked according to our model. Further, our ranking methodology is capable of identifying the highly informative sources much earlier than the widely-used baselines. Our model also shows its potential as an apparatus to analyze events at micro and macro scales. Data for the research is collected from various blogging platforms like blogspot, livejournal, wordpress, typepad, etc. and will be made publicly available for researchers.

4.1 Introduction

Social media has brought a paradigm shift in the way people share information and communicate. Social media played an important role in mobilizing events such as, 'The Arab Spring', 'Occupy Wall Street', 'Sandy relief efforts', 'London Riots','The Spanish Revolution', among others. This led to a surge in citizen journalism all over the world, encouraging transnational participation. Thus, social media serves as a parallel, yet distinct source of information about real-life events along with the mainstream media space [7].

The mainstream media sources often gloss over the intricate details while covering a real-life event. They are often biased, regulated by the government, and may not portray the true picture of an event [8]. While, social media sources like blogs often contain unbiased, uninhibited, and unedited opinions from people. Blogs have been accepted as more credible sources of information over mainstream media sources by the weblog users [9]. Thus the sources, which are obtained from social media could potentially provide a rather 'closer' or an "on-ground" view of the events with novel information. The "on-ground"" information gleaned from the social media affords opportunities to study various online social phenomenon from methodological and theoretical perspectives including, social movements, crowdsourcing, citizen journalism, collective behavior, collective action [10–12], and more.

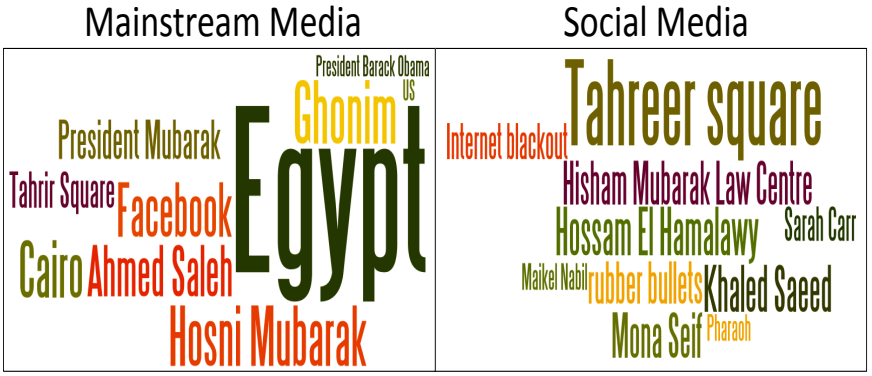


FIGURE 4.1: Top 10 entities from mainstream media and blogs.

Motivation: An initial analysis of the top 10 entities obtained from the top 10 search results related to "Egyptian Revolution" from two mainstream media channels (BBC and CNN), and from blogs during the time of the revolution is shown in Fig 4.1. The top entities from the mainstream media channels are generic and are quiet obvious for the event. In contrast, the top entities from the blogs are very specific to the event. The activists like 'Mona Seif', 'Sarah Carr', 'Maikel Nabil' and 'Hosam El Hamalwy' were very closely involved, and were responsible for mobilizing the event. The entities

like 'Internet Blackout' and 'Khaleed Saeed' were central to the event. Moreover, the presence of entities like 'Facebook' and 'Ghonim' (who was responsible for spreading the event in Facebook) among the top mainstream media entities also indicates the significance of social media in the event.

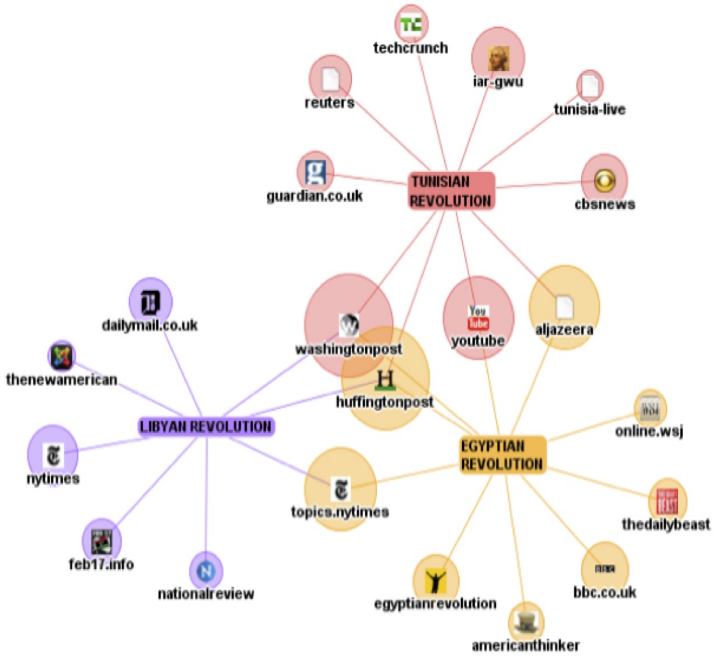


FIGURE 4.2: Top 10 Google search results for "Egyptian Revolution", "Libyan Revolution", and "Tunisian Revolution", visualized using TouchGraph.

Due to the power law distribution of the Internet [13], and the present search engine technology, the 'Short Head' is generally dominated by the mainstream media websites. As illustrated in Fig 4.2 the top 10 search results for "Egyptian Revolution", "Libyan Revolution", and "Tunisian Revolution" by Google, visualized using Touchgraph[1], retrieved mainstream media sources. Consequently, the social media sites get buried in the Long Tail [14] as shown in Fig 4.3. However sources from the social media channels, act as hubs of specific information about real-life events [15]. Thus, a person interested to analyze an event may miss out the novel and specific information available in social media by relying on the top results from the popular search engines. Moreover, in the words of Chris Anderson [16], *"With an estimated 15 million bloggers out there, the odds that a few will have something important and insightful to say are good and getting better."* This motivated us to look for techniques in this paper, that would help in identifying these otherwise buried sources providing highly specific information related to an event.

Challenges: Identifying highly informative 'specific' sources and 'close' entities related to a real-life event from social media entails various challenges as follows,

[1]http://touchgraph.com

- **Sparsity of sources:** Enormous population of the sparsely linked Long Tail social media sources
- **Quality assessment dilemma:** The entities (person, organization, place, etc.) mentioned in the sources act as the atomic units of information. Sources which are 'specific' to an event must contain entities 'closer' or highly relevant to the event. On the other hand, such 'close' entities can be obtained from the 'specific' sources. This presents a dilemma in assessing the quality of the sources for event related 'specific' information content, and makes it a nontrivial task.
- **Entity extraction:** It is also a challenge, to accurately extract the entities from the social media sources, which are mostly unstructured and have colloquial content.
- **Lack of evaluation measures:** Conventional information retrieval based evaluation measures help in identifying the most relevant and authoritative sources, however, these sources may not be the most novel or offer specific information. Therefore, new evaluation measures are required to estimate the performance of our work.

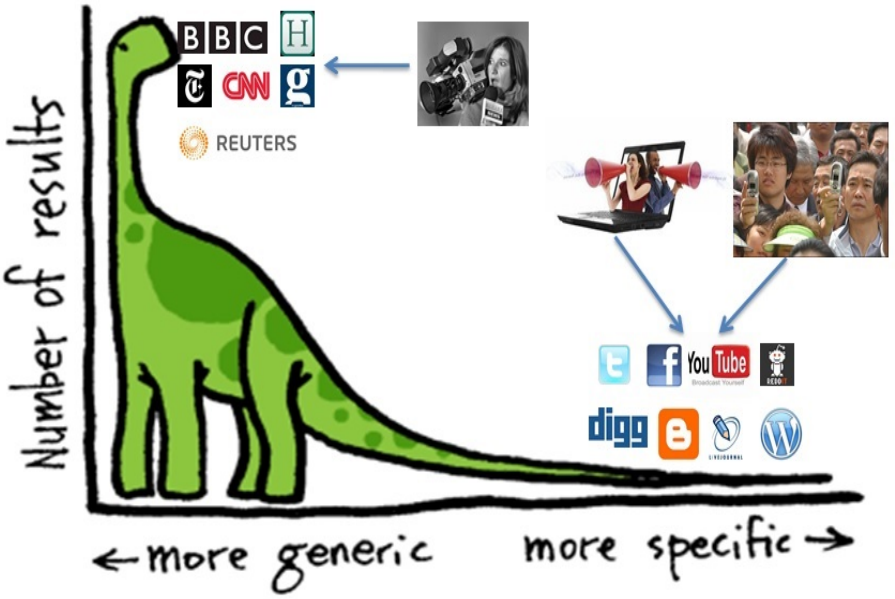


FIGURE 4.3: Short Head Vs Long Tail media sources.

Contributions: We make the following contributions,

- **Methodology:** A methodology based on the principle of mutual reinforcement, that helps in identifying highly 'specific' sources and 'close' entities from social

media, and their relationships (Section 4.4.2). It ranks the sources and the entities based on their 'specific' information content and how 'close' they are with respect to a set of events.

- **Evaluation Strategy:** Present the methodology, along with an objective evaluation strategy to validate our findings (Section 4.6.3).
- **Experiment on sources related to real-life events:** Perform our experiments on sources and entities related to the events: 'Egyptian Revolution', 'Libyan Revolution', and 'Tunisian Revolution' (Section 4.6). However, the work is extendible to other types of events.
- **Event analysis:** Explore the utility of such a model in analyzing events (Section 4.7) and conclude the work with future directions (Section 4.8).

Next, we present the related work and compare and contrast these with the proposed approach, highlighting our contributions to the literature.

4.2 Related Work

In this section, we discuss about some of the previous research relevant to our work. We present studies on analyzing real-life events, identifying quality sources related to them and in general from the web.

User-generated data from various social media platforms, related to real-life events, have been studied to perform wide range of analysis. Socio-political inferences were drawn by studying sentiments and opinions of people towards public and political events from Twitter [17], as well as blogs [18]. Twitter has been extensively used as a source for analyzing information circulated during natural disasters and crisis situations [19, 20]. Tweets related to events have been extracted, summarized and visualized, in order to have a deeper understanding of the events [21, 22].

Our work is different from all such works, and would help in analyzing events from sources and entities, which are highly specific to an event along with the generic ones.

Due to huge number of informal sources in social media it is a difficult task to identify high quality sources related to the real-life events. Researchers have built semantic web models for efficient retrieval of event specific media sources [23]. Event related contents have been found leveraging the tagging and location information associated with the photos shared in Flickr [24]. *Becker et al* [25], studied how to identify events and high quality sources related to them from Twitter. In order to identify the genuine sources

of information, credibility and trustworthiness of event related information were studied from Twitter [26]. New methods were investigated for filtering and assessing the verity of sources obtained from social media by journalists [27].

The work presented in this paper, finds quality sources related to an event from social media, in terms of the 'specific' information content of the source, and is quiet different from all such works.

Several methods have been developed in the past for identifying and ranking quality sources from the web. PageRank [28] took advantage of the link structure of the web for ranking web pages. It was further improved for making it sensitive to topic based search [29]. Graph based approaches were used for modeling documents and a set of documents as weighted text graphs, and for computing relative importance of textual units for Natural Language Processing [**?**]. Mutual reinforcement principle was used for identifying Hubs and Authorities from a subset of web pages [30].

Our work is distinct from all such works. The methodology of our work utilizes the relationship between sources and entities related to a real-life event. It then builds upon the principle of mutual reinforcement and modifies it to a evolutionary system for finding 'specific' sources and 'close' entities w.r.t an event (Subection 4.4.2). A comparative analysis with the conventional mutual reinforcement demonstrates a faster convergence and better performance of our evolutionary model (Subsection 4.6.2). Our framework also has the potential to be used as an apparatus for studying events in terms of the specific sources and close entities identified and ranked by our proposed methodology (Section 4.7). Next, we present the formal problem definition.

4.3 Problem Definition

The number of sources related to an event in social media is overwhelming. All these sources may not provide useful information and needs to be processed in order to identify the valuable sources providing specific information about the concerned event. Provided we have a set of events, a set of sources, and a set of entities related to each of these events, we need to rank these sources and entities from the most specific to the most generic ones, based on their information content.

Event: *We define an event to be a real-world incident, occurring at any place at any time or over a certain period of time.*

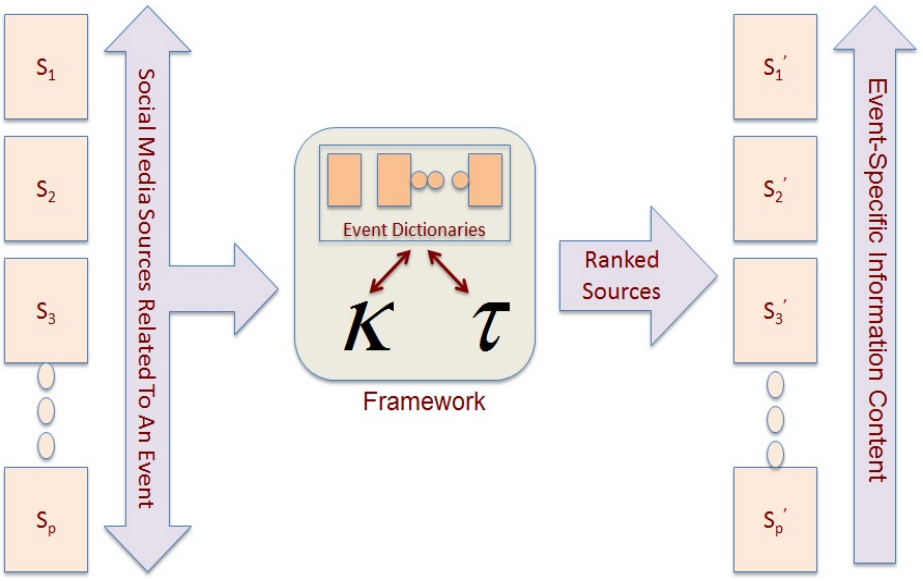


FIGURE 4.4: Black box view of the problem.

Specificity and Closeness: *Given a finite set of events ξ, we take an event $E_j \in \xi$ such that, $1 \leq j \leq |\ \xi\ |$, a set of 'p' sources denoted by ϕ_{E_j}, and a set of 'q' entities denoted by σ_{E_j}, related to the event E_j. We define two functions κ (specificity) and τ (closeness) such that:*

$$\kappa : S_i \rightarrow [0, 1] \tag{4.1}$$

$$\tau : e_i \rightarrow [0, 1] \tag{4.2}$$

where, $S_i (\in \phi_{E_j})$, is the i^{th} source, and $e_i (\in \sigma_{E_j})$ is the i^{th} entity, so that we can get two ordered sets (φ_{E_j} and ς_{E_j}) for the set of sources in ϕ_{E_j} and entities in σ_{E_j}, such that:

$$\varphi_{E_j} = \{S_1, ..., S_i, S_j, ..., S_p \mid \kappa(S_i) \geq \kappa(S_j), i < j\} \tag{4.3}$$

$$\varsigma_{E_j} = \{e_1, ..., e_i, e_j, ..., e_q \mid \tau(e_i) \geq \tau(e_j), i < j\} \tag{4.4}$$

φ_{E_j} is ordered in decreasing order of how 'specific' S_i is w.r.t E_j. ς_{E_j} is ordered in decreasing order of how 'close' e_i is w.r.t E_j. A black-box view of the problem is shown in Fig 4.4.

4.4 Methodology

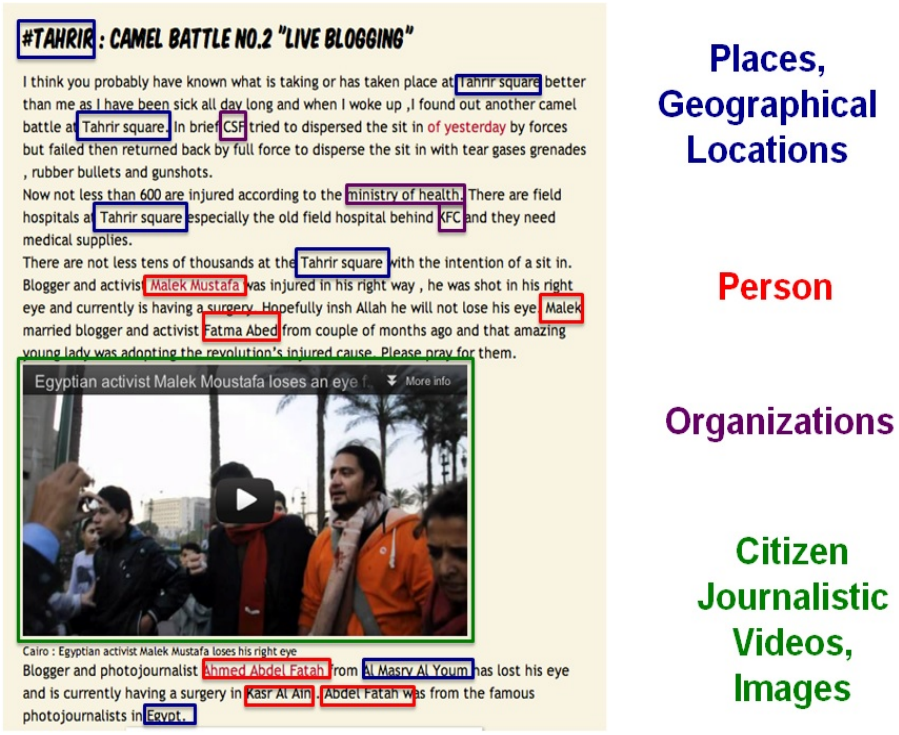


FIGURE 4.5: Entities associated with a social media source.

A real-life event is characterized by a distinct set of close entities (persons, places, organizations, etc.) along with generic ones. The entities act as the basic units of information in these sources as shown in Fig 4.5. Intuitively, specific sources would contain closer entities and one is likely to find closer entities in more specific sources. The relation between specific sources and close entities could then be modeled following the Mutual Reinforcement Principle, which forms the basis of our methodology.

> An entity should have high 'closeness' score if it appears in many sources with high 'specificity' scores while a source should have a high 'specificity' score if it contains many entities with high 'closeness' scores.

In essence the principle states that the 'closeness' score of an entity is determined by the 'specificity' scores of the sources it appears in, and the 'specificity' score of a source is determined by the 'closeness' scores of the entities it contain. The proposed methodology extends the basic Mutual Reinforcement Principle to consider the evolving knowledge learned about an event. However, the model requires an apriori or seed knowledge about an event, which is provided in terms of an event profile or an event dictionary. Next, we discuss the construction of event dictionaries.

4.4.1 Event Dictionaries

Each event E_j is profiled by constructing an event dictionary (σ_{E_j}). In order to calculate specificity of a source w.r.t an event, we need to start with an initial set of close entities. At the same time, these close entities are better acquired from the specific sources. To solve this dilemma, we construct event dictionaries, from independent sources which are completely separate from the sources (ϕ_{E_j}) that need to be ranked.

Formulation of initial closeness scores: We calculate the 'closeness' score ($\tau(e_i)_{E_j}$) of each entity (e_i) for event E_j in order to construct the event dictionaries, by using equations 4.5 and 4.6 based on tf-idf measure [31] , from the information retrieval literature. Let $E_j \in \xi$, be the j^{th} event, and 'e_i' be the i^{th} entity extracted from the set of sources selected for constructing the event dictionaries. If the term $f(e_i, E_j)$ denotes the frequency of occurrence of the entity 'e_i' in the set of sources for the event E_j, and $IE_jf(e_i)$ denotes the inverse event frequency for the entity 'e_i' then closeness score ($\tau(e_i)_{E_j}$) of an entity e_i w.r.t the event E_j is defined as,

$$\tau(e_i)_{E_j} = e_i f\_IE_j f = f(e_i, E_j) * IE_j f(e_i) \tag{4.5}$$

$$IE_j f(e_i) = log(\frac{\mid \xi \mid}{\mid E_j \in \xi : e_i \in E_j \mid}) \tag{4.6}$$

and, $\mid E_j \in \xi : e_i \in E_j \mid$ refers to the number of events in which the entity e_i occurs. Since we extract the entities from the sources related to the events, we cannot have an entity that does not belong to any of the events. Therefore, we always have $\mid E_j \in \xi : e_i \in E_j \mid > 0$.

We get $\mid \xi \mid$ number of event dictionaries, each corresponding to an event. Following steps are taken to construct the event dictionaries:

1. ***Entity Extraction:*** Entities are extracted from all the sources collected from GlobalVoices[2] as explained in Section 4.5, using AlchemyAPI[3] and their corresponding $\tau(e_i)_{E_j}$ values are calculated using equation 4.5. We choose GlobalVoices for obtaining the seed sources for constructing the initial event dictionaries, as it is a portal where bloggers and translators work together to make reports of various real-life events, from blogs and citizen media everywhere.This makes it a reliable

[2]http://globalvoicesonline.org
[3]http://alchemyapi.com

source for finding specific information content from social media. Due to colloquial nature of the sources as discussed in the challenges, some of the entities occur in several forms. For example, the entity 'Tahrir Square' occur as 'Tahreer', 'El-Tahrir', etc. We resolve such multiple representation of the same entity by applying pattern matching[4]. Given two entities represented as strings we accept them to be the same if their patterns match by 80% or more. We would like to use the standard entity resolution algorithms in our future work.

2. ***Closeness Score Computation:*** For each event E_j, we calculate $\tau(e_i)_{E_j}$ scores for the set of entities for that event using equations 4.5 and 4.6. An entity may occur in multiple events and hence can be present in multiple event dictionaries with different $\tau(e_i)_{E_j}$ scores.

3. ***Ranking:*** The higher the $\tau(e_i)_{E_j}$ score of an entity the closer it is to the event. The entities are then ranked according to the descending $\tau(e_i)_{E_j}$ scores.

4. ***Normalization:*** Since the range of closeness scores are different for each event, we normalize $\tau(e_i)_{E_j}$ scores w.r.t an event between 0 and 1. The normalization enables an assessment of relative closeness of an entity across multiple events.

The dictionaries thus obtained from the above mentioned procedure are static and serve as a good source of apriori knowledge about the event. However, as we discover new knowledge from specific sources, it is desirable to update the event dictionaries. However, the method applied for constructing the initial event dictionaries require a set of events. This is a drawback of the current method and we plan to improve it in a future work. Next, we discuss how the dictionaries help in identifying specific sources, which in turn help in improving the dictionary.

4.4.2 Mutually Reinforcing Sources and Entities

Given an event $E_j \in \xi$, a set of sources (ϕ_{E_j}) and entities (σ_{E_j}), related to the event, we define two column vectors: ***'Specificity'*** ($\kappa_{\mathbf{E_j}}$) and ***'Closeness'*** ($\tau_{\mathbf{E_j}}$).

$$\kappa_{\mathbf{E_j}} =< \kappa(S_1)_{E_j}, \kappa(S_2)_{E_j}, ..., \kappa(S_p)_{E_j} >^T \tag{4.7}$$

$$\tau_{\mathbf{E_j}} =< \tau(e_1)_{E_j}, \tau(e_2)_{E_j}, ..., \tau(e_q)_{E_j} >^T \tag{4.8}$$

[4]http://docs.python.org/2/library/difflib.html

where, $\kappa(S_i)_{E_j}$ ($\in range(\kappa)$, from equation 4.1) represents the *'specificity'* score of i^{th} source $S_i(\in \phi_{E_j})$, for $1 \leq i \leq p$ and $\tau(e_i)_{E_j}$ ($\in range(\tau)$, from equation 4.2) represents the *'closeness'* score of i^{th} entity $e_i(\in \sigma_{E_j})$, for $1 \leq i \leq q$. Each source S_i may contain related as well as unrelated information about various events. If we consider the set of events ξ, then each $\kappa(S_i)_{E_j}$ is itself a vector of 'specificity' values of the source S_i w.r.t the events ($\in E_j$) as expressed in equation 4.9.

$$\kappa(S_i)_{E_j} =< \kappa(S_i)_{E_1}, \kappa(S_i)_{E_2}, ..., \kappa(S_i)_{E_{|\xi|}} > \tag{4.9}$$

Similarly, each entity e_i may be related to various events. If we consider the set of events ξ, then each $\tau(e_i)_{E_j}$ is itself a vector of 'closeness' values of the entity e_i w.r.t the events ($\in E_j$) as expressed in equation 4.10.

$$\tau(e_i)_{E_j} =< \tau(e_i)_{E_1}, \tau(e_i)_{E_2}, ..., \tau(e_i)_{E_{|\xi|}} > \tag{4.10}$$

However, while representing $\kappa(S_i)_{E_j}$ and $\tau(e_i)_{E_j}$ as an element of the vectors $\kappa_{\mathbf{E_j}}$ and $\tau_{\mathbf{E_j}}$, respectively, we only choose the entry for the j^{th} event under consideration.

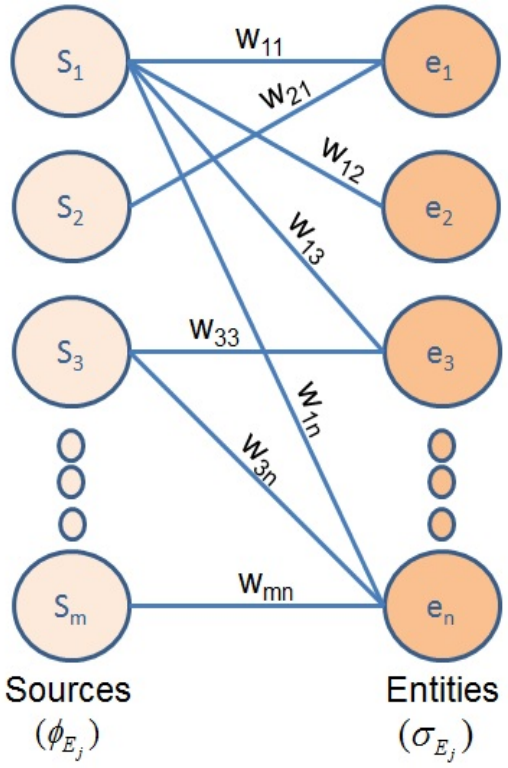


FIGURE 4.6: Bipartite graph G representing the mutual relationship between the sources and the entities.

We construct a bipartite graph $G = (V, U)$ (Figure 4.6) representing the mutual relationship between the sources and the entities, where $V \in \phi_{E_j}, \sigma_{E_j}$, is the set of vertices for G, and U is the set of undirected edges. The sources without entities are discarded during this process.

The presence of an entity in a source is not sufficient to determine its specificity. In order to express the specificity of a source w.r.t an event, we need to consider the closeness value of the entities present in the source. Given the closeness $\tau(e_n)_{E_j}$ of an entity e_n w.r.t an event E_j, obtained from the event dictionary, a weight w_{mn} is assigned to the edges of the graph, which expresses the magnitude by which an entity is related to a source S_m.

The significance of an entity e_n in a source S_m is expressed as,

$$\frac{f(e_n, S_m)}{\sum_{n=0}^{q} f(e_n, S_m)} \tag{4.11}$$

where, $f(e_n, S_m)$ is the frequency of occurrence of the entity e_n in the source S_m. Therefore mathematically,

$$w_{mn} = \frac{\tau(e_n)_{E_j} * f(e_n, S_m)}{\sum_{n=0}^{q} f(e_n, S_m)} \tag{4.12}$$

The adjacency matrix of the bipartite graph G is denoted by L, and is defined as follows:

$$L_{mn} = \begin{cases} w_{mn} & if\ (m,n) \in U \\ \\ 0 & otherwise \end{cases}$$

Following the Mutual Reinforcement Principle the relationships between specificity scores of sources and closeness scores of entities for event E_j can be denoted as follows,

$$\kappa_{\mathbf{E_j}} = \mathbf{L}\tau_{\mathbf{E_j}} \tag{4.13}$$

$$\tau_{\mathbf{E_j}} = \mathbf{L^T}\kappa_{\mathbf{E_j}} \tag{4.14}$$

Substituting the values for $\kappa_{\mathbf{E_j}}$ and $\tau_{\mathbf{E_j}}$, we derive the following equations,

$$\kappa_{\mathbf{E_j}} = \mathbf{L}\mathbf{L^T}\kappa_{\mathbf{E_j}} \tag{4.15}$$

Input: Set of sources ϕ_{E_j}, set of entities σ_{E_j} from the event dictionary, threshold for convergence μ.
Output: Ordered set of sources φ_{E_j}, ranked according to their 'specificity' scores w.r.t E_j, and, ordered set of entities ς_{E_j}, ranked according to their 'closeness' scores w.r.t E_j.

Initialize $\kappa_{\mathbf{E_{j(0)}}} \leftarrow \tau_{\mathbf{E_{j(0)}}} \leftarrow < 1, 1, ..., 1 >$;
Initialize $k \leftarrow 1$;
repeat
  Construct matrices $\mathbf{L_{k-1}}$ and $\mathbf{L^T_{k-1}}$;
  Calculate matrix products $M \leftarrow \mathbf{L_{k-1}L^T_{k-1}}$ and $M'' \leftarrow \mathbf{L^T_{k-1}L_{k-1}}$;
  Convert M and M'' into stochastic matrices
  $M_{stochastic} \leftarrow M$ *and* $M''_{stochastic} \leftarrow M''$;
  $\kappa_{\mathbf{E_j}}(\mathbf{k}) \leftarrow M_{stochastic}\kappa_{\mathbf{E_j}}(\mathbf{k-1})$;
  $\tau_{\mathbf{E_j}}(\mathbf{k}) \leftarrow M''_{stochastic}\tau_{\mathbf{E_j}}(\mathbf{k-1})$;
  normalize $\kappa_{\mathbf{E_j}}(\mathbf{k}) \leftarrow \frac{\kappa_{\mathbf{E_j}}(\mathbf{k})}{||\kappa_{\mathbf{E_j}}(\mathbf{k})||_1}$;
  normalize $\tau_{\mathbf{E_j}}(\mathbf{k}) \leftarrow \frac{\tau_{\mathbf{E_j}}(\mathbf{k})}{||\tau_{\mathbf{E_j}}(\mathbf{k})||_1}$;
  $k \leftarrow k + 1$;
  Update σ_{E_j} ;
until $|| \kappa_{\mathbf{E_j}}(\mathbf{k}) - \kappa_{\mathbf{E_j}}(\mathbf{k-1}) ||_1 < \mu$ *and* $|| \tau_{\mathbf{E_j}}(\mathbf{k}) - \tau_{\mathbf{E_j}}(\mathbf{k-1}) ||_1 < \mu$;
Reverse sort $\kappa_{\mathbf{E_j}}(\mathbf{k})$,$\tau_{\mathbf{E_j}}(\mathbf{k})$;
return $\kappa_{\mathbf{E_j}}(\mathbf{k})$,$\tau_{\mathbf{E_j}}(\mathbf{k})$;

FIGURE 4.7: Algorithm for calculating 'specificity' and 'closeness'.

$$\tau_{\mathbf{E_j}} = \mathbf{L^T L}\tau_{\mathbf{E_j}} \tag{4.16}$$

Equations 4.15 and 4.16 are characteristic equations of an eigensystem, where the solutions to $\kappa_{\mathbf{E_j}}$ and $\tau_{\mathbf{E_j}}$ are the respective eigen vectors with the corresponding eigenvalue of 1.

To emphasize the relationship between the sources and the entities, we make a major contribution by modifying the way the equations 4.15 and 4.16 are solved. We make the matrices $\mathbf{LL^T}$ and $\mathbf{L^T L}$ evolutionary while solving the equations. Since each of the equations is a circular definition, the final specificity and closeness scores are computed using the power iteration method [32]. Each iteration improves specificity and closeness scores reflecting their mutual relationship. As we move towards getting the specific sources and close entities in each iteration, we update the weights (w_{mn}) assigned to the edges between the sources and the entities by the newly calculated closeness scores for the entities. This results in renewed reinforcement of the relationship at every iteration by getting closer entities from better sources and vice-versa. This essentially helps the model incorporate the newly discovered knowledge about the events. More precisely, the improved understanding of the relationship between the source and the entities vis-a-vis an event is incorporated into the model.

The updation of the edge weights and the matrices with k^{th} iteration is represented as follows,

$$w_{mn(k)} = \frac{\tau(e_{n(\mathbf{k-1})})_{E_j} * f(e_n, S_m)}{\sum_{n=0}^{q} f(e_n, S_m)} \tag{4.17}$$

$$L_{mn(k)} = \begin{cases} w_{mn(k)} & if\ (m,n) \in U \\ 0 & otherwise \end{cases}$$

where, $L_{mn(k)}$ represents the adjacency matrix for graph G, and $w_{mn(k)}$ denotes the edge weight for the edge between m^{th} source and n^{th} entity at the k^{th} iteration. $\tau(e_{n(\mathbf{k-1})})_{E_j}$ represents the closeness score of the entity e_n w.r.t the event $E_j (\in \xi)$, obtained from the evolving event dictionary for event E_j at $(k-1)^{th}$ iteration.

If, $\kappa_{\mathbf{E_j}}(\mathbf{k})$ and $\tau_{\mathbf{E_j}}(\mathbf{k})$ be the specificity and the closeness scores, at the k^{th} iteration, the iterative process for generating the final solution are,

$$\kappa_{\mathbf{E_j}}(\mathbf{k}) = \mathbf{L_{k-1}} \mathbf{L_{k-1}^T} \kappa_{\mathbf{E_j}}(\mathbf{k-1}) \tag{4.18}$$

$$\tau_{\mathbf{E_j}}(\mathbf{k}) = \mathbf{L_{k-1}^T} \mathbf{L_{k-1}} \tau_{\mathbf{E_j}}(\mathbf{k-1}) \tag{4.19}$$

In order to get 1 as the largest eigenvalue and, $\kappa_{\mathbf{E_j}}$ and $\tau_{\mathbf{E_j}}$ as the principal eigen vectors, the matrices $\mathbf{L_{k-1}} \mathbf{L_{k-1}^T}$ and $\mathbf{L_{k-1}^T} \mathbf{L_{k-1}}$ needs to be stochastic and irreducible [33] at every step of our evolutionary process. In the present case, since the graph G is a bipartite graph, matrices $\mathbf{L_{k-1}} \mathbf{L_{k-1}^T}$ and $\mathbf{L_{k-1}^T} \mathbf{L_{k-1}}$ are already irreducible.

In order to make the matrices $\mathbf{L_{k-1}} \mathbf{L_{k-1}^T}$ and $\mathbf{L_{k-1}^T} \mathbf{L_{k-1}}$ stochastic, we take the following steps at each iteration,

- Dividing the non-zero entries of the matrices $\mathbf{L_{k-1}} \mathbf{L_{k-1}^T}$ and $\mathbf{L_{k-1}^T} \mathbf{L_{k-1}}$ by the summation of all the entries in a row.
- Assigning $1/n$ to the zero entries of $\mathbf{L_{k-1}} \mathbf{L_{k-1}^T}$ and $1/m$ to the zero entries of $\mathbf{L_{k-1}^T} \mathbf{L_{k-1}}$, respectively.

The whole process is presented as an algorithm as shown in Fig 4.7.

We also perform our study using conventional binary static matrices represented as follows,

$$L_{mn} = \begin{cases} 1 & if\ (m,n) \in U \\ \\ 0 & otherwise \end{cases}$$

The proposed evolutionary model outperforms the static model as validated by the results discussed in Section 4.6.

4.5 Data Collection

Motivation behind source selection: For many people blogs have become popular social media sources for satisfying interpersonal communication needs. Blogs act as a platform for masses to share their likes and dislikes, voice their opinions, provide suggestions and report news. Over the years blogging has matured from personal diaries to citizen journalistic sources providing live coverage of events beyond the professional newsrooms. Often mainstream media rely on blogs for reporting first-hand accounts of an event [34]. Other social media platforms like microblogs, social networks etc., also promote such activities. But, these platforms have very little scope to elaborately discuss about the events due to the limitations in the length of content allowed to be posted. However, these alternative platforms act as good sources for studying and tracking dissemination of information during real-life events. This motivated us to perform our experiments on sources collected from the blogging platforms instead of other social media websites.

TABLE 4.1: Details of Data Collected.

Service Used	Event	Number of Blog Posts
GlobalVoices	Egyptian Revolution	234
	Libyan Revolution	86
	Tunisian Revolution	77
Google Blogger	Egyptian Revolution	579
	Libyan Revolution	600
	Tunisian Revolution	484
Icerocket Blog Search	Egyptian Revolution	5900
	Libyan Revolution	2198
	Tunisian Revolution	1220

Sources: Blog posts from GlobalVoices, Blogger[5] and Icerocket Blog Search[6] respectively, are collected for the study. The details of the dataset used is given in Table 4.1. The dataset includes 11,378 blog posts from various blogging platforms like blogspot.com, wordpress.com, livejournal.com, typepad.com, etc. We also filter out the non-english blogs. The data from GlobalVoices is used for constructing event dictionaries (σ_{E_j}), as explained in Section 4.4. We collect blog posts related to the three events from Blogger using Google Search, and from other blogging platforms using Icerocket blog search. We perform our experiments on the sources (ϕ_{E_j}) retrieved by the search engines due to the lack of ground truth and take the sources along with the ranks assigned to them by the search engines as our baseline (explained in Subsection 4.6.3) The collected blog posts are parsed for extracting various information. However, we use the following information for our study: *URL* of the blog and blogpost), *blog text*, *entities*, *language*, and *rank* of the post in the respective search engines used for collecting it. We use AlchemyAPI in order to extract entities. These datasets would be made available on request.

4.6 Experiment and Analysis

In this section, we describe the experiments performed. First, we discuss the experimental setup, followed by the comparative analysis between the proposed evolutionary mutual reinforcement model and conventional mutual reinforcement model. We then, introduce a novel evaluation strategy comparing the proposed model with two baseline models.

4.6.1 Experimental Setup

The methodology discussed earlier is implemented on the collected datasets. We take the following steps in order to perform the experiment,

- ***Constructing the Event Dictionaries:*** We take $\xi = \{$"Egyptian Revolution", "Libyan Revolution", "Tunisian Revolution"$\}$ as our set of events. We construct the event dictionaries (σ_{E_j}) by using the sources from GlobalVoices (explained in Event Dictionary subsection).
- ***Implementing the Proposed Evolutionary Mutual Reinforcement Model:*** The algorithm, as presented in Figure 4.7, is implemented on the set of sources

[5]http://blogger.com
[6]http://icerocket.com

(ϕ_{E_j}) from Blogger and Icerocket related to each event respectively, and the set of entities (σ_{E_j}) from the event dictionary corresponding to each event E_j. The threshold value for convergence μ is set to 1e-08.

- ***Obtaining Specific Sources and Close Entities:*** With the termination of the algorithm we get a ranked set of sources (φ_{E_j}) ordered in terms of their specific information content and entities (ς_{E_j}) ordered in terms of how closely related they are to the event.

- ***Conventional Mutual Reinforcement Model Approach:*** We also implement the conventional mutual reinforcement model on ϕ_{E_j} and σ_{E_j} without considering the evolving matrices (explained in Subsection 4.4.2).

4.6.2 Comparing Conventional and Evolutionary Mutual Reinforcement Models

In order to compare the efficiency of the proposed evolutionary mutual reinforcement model with the conventional mutual reinforcement model we combine the sources collected for an event E_j from Google search and Icerocket search. After that we run the proposed evolutionary mutual reinforcement model and conventional mutual reinforcement model on the combined set of sources and event dictionary (σ_{E_j}) for each event. The number of iterations taken by the power iteration method to converge in each case is analyzed in Figure 4.8. It is observed that the number of iterations taken by the power iteration method is lesser, when we employ the evolutionary mutual reinforcement model. Hence, we observe a marked improvement in the performance of our evolutionary mutual reinforcement model over the conventional static mutual reinforcement model.

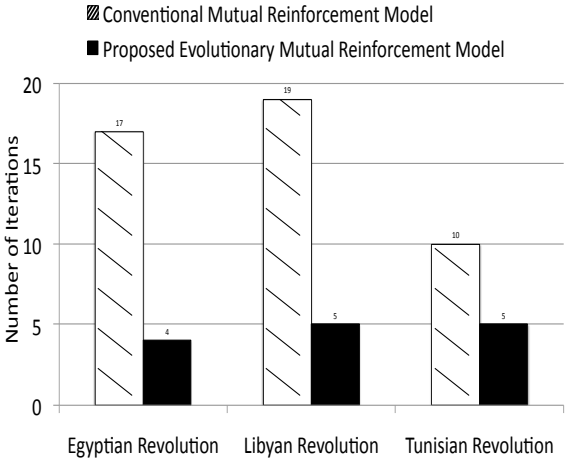


FIGURE 4.8: Comparison of number of iterations taken by the power iteration method to converge, for the set of sources related to events in ξ, with the proposed evolutionary mutual reinforcement model and the conventional static mutual reinforcement model.

Explanation for the improvement: The lesser number of iterations in the proposed evolutionary mutual reinforcement model can be explained by the introduction of the evolutionary weights assigned to the relationship between the sources and the entities. The improved closeness scores ($\tau(e_i, E_j)$) at each iteration of the evolutionary model and the renewed reinforcement of the relationship between the entities and the sources decreases the number of iterations taken by the proposed evolutionary model to converge in comparison to the static conventional model. Next, we compare the performance of the proposed evolutionary model with the baselines and the conventional mutual reinforcement model in terms of how quickly the models help in identifying valuable information about the events.

Egyptian Revolution

Specificity Based Ranking	Google Search Ranking
1	59
2	286
3	400
4	277
5	55
6	202
7	6
8	9
9	313
10	374

Libyan Revolution

Specificity Based Ranking	Google Search Ranking
1	13
2	329
3	9
4	194
5	24
6	311
7	364
8	204
9	374
10	184

Tunisian Revolution

Specificity Based Ranking	Google Search Ranking
1	162
2	40
3	420
4	459
5	72
6	181
7	152
8	440
9	99
10	174

Egyptian Revolution

Specificity Based Ranking	Icerocket Blog Search Ranking
1	75216
2	10607
3	53924
4	56604
5	9831
6	25790
7	1
8	99925
9	94614
10	53924

Libyan Revolution

Specificity Based Ranking	Icerocket Blog Search Ranking
1	47276
2	11751
3	4900
4	22501
5	4
6	11040
7	43520
8	11751
9	41631
10	18271

Tunisian Revolution

Specificity Based Ranking	Icerocket Blog Search Ranking
1	9713
2	42985
3	36335
4	3843
5	46784
6	42645
7	99
8	1
9	63141
10	42645

FIGURE 4.9: Rankings of the sources from Google Blogger and Icerocket based on 'specificity' (κ) values obtained from our model and the rankings assigned by Google Search and Icerocket Blog Search.

4.6.3 Baseline Comparisons

Selecting the Baselines: Due to lack of benchmark datasets we use the search results obtained from Google and Icerocket Blog Search as baselines for validation. Standard information retrieval measures for evaluation (DCG, NDCG, MAP, MAP@10) could not be used due to the absence of ground truth. We also consider the sources ranked according to the conventional static mutual reinforcement model and show the effectiveness

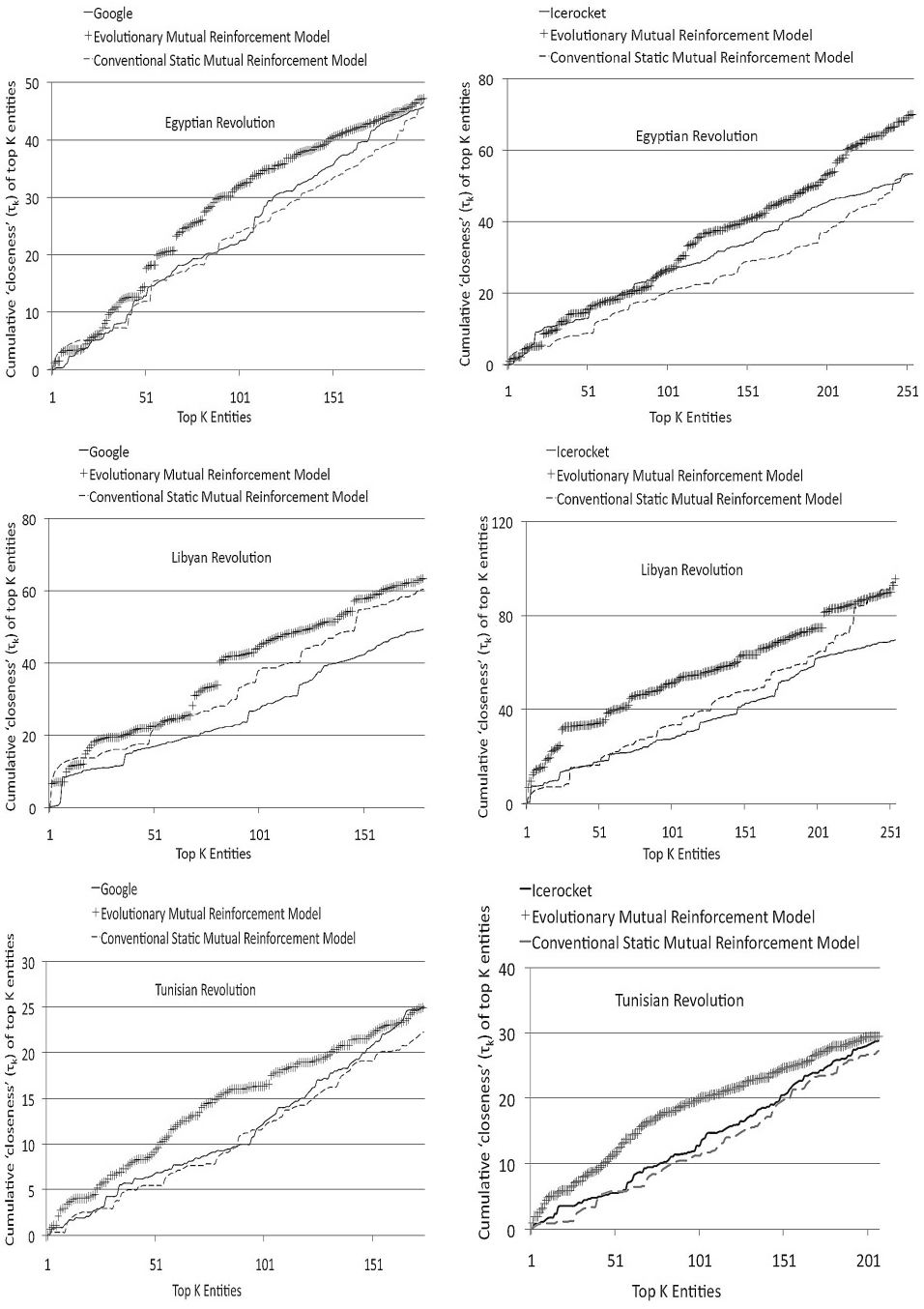


FIGURE 4.10: Validating specific sources obtained from our model.

of our model in quickly gaining information about an event. We further analyze the ranking of the top K specific sources as identified by our methodology and observe the difference in their rankings as assigned by the search engines and as assigned by our model. Figure 4.9 shows ranks assigned by Google and Icerocket for the top 10 sources for each event, ranked according to our framework. We conclude, that our framework could identify the sources that often gets buried in the Long Tail and has the potential for presenting valuable information about the event. Next, we propose a novel strategy to demonstrate the effectiveness of the specific sources ranked by our evolutionary mutual reinforcement model in identifying highly informative sources.

Rationale Behind The Strategy: Search engines are designed to give the most relevant sources containing the close entities, related to a query for a given event. As these close entities have a very high probability to be associated with the event, we expect to gain valuable information about the event from these entities making the highly ranked sources very specific to the event. We use this notion to propose a novel evaluation strategy, showing that when sources related to an event are ranked according to our model, they provide more valuable information than the ranking order given by the search engines and the conventional mutual reinforcement model for the same set of sources.

Implementation of The Strategy: We compare closeness ($\tau(e_i)_{E_j}$) values of the entities obtained from the sources ranked according to the search engines, the conventional model and our model respectively. Following steps are taken,

- ***Preparing the Ranked Lists of Entities:*** From the three differently ranked lists (search engine, our evolutionary model and conventional static model), each source is visited and the entities are extracted from them. The '$\tau(e_i)_{E_j}$' values are assigned to these entities by referring to the respective final event dictionary (ς_{E_j}) and they are ranked in descending order.

- ***Measuring the Information Gain:*** As we traverse the three list of sources, we obtain three list of same entities, arranged in different orders depending upon the ranking of the sources. In order to show the comparison in the gain of information from the three lists we take the top 'K' entities from each list and calculate the sum of their '$\tau(e_i)_{E_j}$' values and plot them against the value of 'K' in Figure 4.10. We start from K=1 and go on increasing its value till the number of entities are exhausted in all the three lists. It is evident from Figures 4.10, that the curves based on specificity ($\kappa(S_i)_{E_j}$) quickly gains over the curves based on the search engines and the conventional model.

- ***Analysis:*** Figure 4.10 shows that information about the event is gained quicker using our model, which could identify specific sources earlier than the search engines as well as the conventional model. We measure the maximum percentage gain of information in each set of sources related to the three events. There is a maximum gain of information (130.4%) in case of sources obtained from Icerocket search related to 'Tunisian Revolution', and a minimum gain of information (25.97%) in case of the sources obtained from Icerocket search related to 'Egyptian Revolution'. Also when the sources are ranked according to our methodology they gain the maximum percentage of information at the 9^{th} source in case of sources obtained from Google related to 'Libyan Revolution'. This in turn implies that the sources ranked higher by our model are more specific than the ones ranked by the search engines and the conventional model. These highly specific sources are also very informative. When presented earlier they also help in learning useful information about the event due to the presence of close entities in them. As we already observed earlier that these sources are often Long Tail sources, we can conclude that Long Tail sources that are ranked lower by the search engines, when identified are often more specific than the highly ranked short head sources.

TABLE 4.2: Top 5 entities in the event specific and the event class dictionaries constructed for the set of events ξ.

Egyptian Revolution Specific Dictionary	Tahrir Square, Egyptian government, Gigi Ibrahim, Alexandria, Wael Abbas.
Libyan Revolution Specific Dictionary	Tripoli, Muammar Al Gaddafi, North Atlantic Treaty Organization, Chad, United Kingdom
Tunisian Revolution Specific Dictionary	Tunisian government, Lin Ben Mhenni, Samir Feriani, Kasbah Square, RCD
Socio-Political event dictionary	Twitter, Iranian Government, Tear gas devices, Facebook, Big Social network

4.7 Further Exploration

We use the proposed framework developed by us as an apparatus to further explore event characteristics and show its utility in analyzing events. We show the potential of the final event dictionaries (ς_{E_j}) obtained for each event E_j for gaining valuable information about the event and to identify the generic entities related to the class of events.

	Frequent (η >= 5)	(η < 5) Not Frequent
(α >= 0.5) Close	Egyptian government, Tahrir Square, Suez, Gigi Ibrahim, Alexandria, Maikel Nabil, NDP, Muslim Brotherhood, Egyptian Army, Wael Ghonim.	Internet Café, Day of Anger, Mariam Arafat, Candice Holdsworth, Dalia Al Marghani, Sherine Tadros, EGP, Bahaa El-Tawil, Mir Hussein Mousavi, Hussein Sharif.
Not Close (α < 0.5)	Anwar al-Sadat, Open Society Institute, Facebook, Professor Rashid Khalidi, dictator, Mohammed Abdel Dayem, mass movement, illegal occupation, Tea Party, Abdel Salam Karmen.	Mainstream media, Oman, Tunis, Carlos Latuff, Ukraine, Sudan, Cuba, Hillary Clinton.

Egyptian Revolution

	Frequent (η >= 5)	(η < 5) Not Frequent
(α >= 0.5) Close	Saif Al Islam Gaddaffi, Pentagon, oil, Mohamed Nabbous, Central Africa, Muammar Gaddafi, Rwanda, Arab League, Ethiopia, Khamis Gaddaffi	North Atlantic Treaty Organization, Iyad El Baghdadi, Libyan State Television, Bent Benghazi, Altarash, Benina airport, Omar Al-Mukhtar, Youssef Al Qaradawi, Eman Al Obaidi, Hamdi Kadri
Not Close (α < 0.5)	Tony Buckingham, Senoussis Brotherhood, Safia Farkash, Cynthia McKinney, United States of America, Canada, popular media, United Africa, Emad Benosman, Hosni Mubarak	Mainstream media, Oman, Tunis, Bahrain, UAE, Saudi Arabia, New York Times, Afghanistan, Ben Ali, Wael Ghonim

Libyan Revolution

	Frequent (η >= 5)	(η < 5) Not Frequent
(α >= 0.5) Close	Ennahda party, Moncef Marzouki, Sidi Bouzid, Tunisian government, Habib Bourguiba, Al Abedine Ben Ali, Muslim Brotherhood, RCD, Mohammed Ghannouchi.	Kasbah Square, France, Slim Amamou, Amira Yahyaoui, Beji Caid Al Sebsi, Mehdi Lamloum, Sami Ben Gharbia, Aziza Othmana Hospital, Samar Dahmash Jarrah, Sami Ben Romdhane
Not Close (α < 0.5)	Michele Alliot-Marie, social networks, facebook, Arabs, Jeffrey Feltman, United States of America, Al-Mahdi, Al-Qaida, Al Jazeera, Kareem Salama	Mainstream media, Oman, Doha, Tunis, UAE, Saudi Arabia, New York Times, Hosni Mubarak, Bill Clinton, Hillary Clinton

Tunisian Revolution

FIGURE 4.11: Different categories of entities for the events.

The entities in the final event dictionaries (ς_{E_j}) are examined further. Based on their frequency $f(e_i, E_j)$ and closeness ($\tau(e_i)_{E_j}$) scores, we categorize the entities into the following categories: a. *Close and Frequent*, b. *Close but Not Frequent*, c. *Not Close but Frequent*, and d. *Neither Close nor Frequent* as shown in Figure 7. We categorize an entity in each event dictionary as 'frequent' if it occurs more than a threshold value η i.e $f(e_i, E_j) \geq \eta$, and 'close' if it has a closeness score more than a threshold value α i.e $\alpha \geq \tau(e_i, E_j)$. After careful manual inspection we decided the values of η and α to be 5 and 0.5 respectively. The values of η and α is same for all the events $E_j \in \xi$. However, different thresholds could be examined in the future though. Each entry in a matrix, consists of top 10 entities that satisfy the thresholds for the corresponding row and column category. We further examine these entities in the context of each of these events and report interesting observations. We present the observations for "Egyptian Revolution", next, however similar observations were made for other events.

4.7.1 Event-specific Popular and Close Entities

We analyze the entities in the *Close and Frequent*, and *Close but Not Frequent* categories. We find that *Close and Frequent* entities are not only frequent but are also closely related to the event. In other words these are very popular entities related to the event. For example, the occurrence of 'Tahrir Square' and 'Egyptian government' in this category, is inevitable, as Tahrir Square is the place where the protest started against the Egyptian Government. These are also the entities that anyone comes to know about from the top search results given by the search engines as well as from mainstream media sources.

On the other hand, *Close but Not Frequent* category primarily consists of entities that add novel and useful insights to the events. The occurrence of 'Internet Cafe' in this category clearly shows how the local people in Egypt used Internet Cafes for accessing various social media websites in order to coordinate and participate in the revolution. It is also the place where 'Khaled Said', a 28 year old computer-wiz was arrested by the Egyptian police. He was brutally tortured to death, triggering the anger among the people of Egypt for a mass uprising against the dictatorial government. January 25, 2011 is also known as 'The Day of Anger' in the Egyptian Revolution, as it is the day that marked the start of a series of protests and riots in Egypt. So the occurrence of the entity, 'Day of Anger' in this category is reasonable.

Drawing upon the differences between the two categories our findings suggests that although the category of *Close and Frequent* entities give a lot of information about an event, it is also necessary to know about the entities of *Close but Not Frequent* category. The entities belonging to *Close but Not Frequent* category provides in-depth information

about the event from the grass root level. The *Close but Not Frequent* entities are likely to be found buried in the Long Tail sources. In order to gain maximum information about an event it is necessary to identify the entities from both the categories. They point to the key persons, places and organizations related to the event. In other words, these entities when present in a source makes it highly specific to the event. Both these categories of entities can prove to be vital sources of information about the event.

4.7.2 Event-specific and Event-class specific dictionaries

Based on the categorization and the analysis conducted in the previous subsection we divide the final event dictionaries (ς_{E_j}) for each event into *event-specific* and *event-class specific* dictionaries. The *event-specific* dictionaries are comprised of the entities categorized as *Close and Frequent*, and *Close but Not Frequent* for each event $E_j \in \xi$ respectively. Whereas, the *event-class specific* dictionary is comprised of the entities found in the *Not Close but Frequent* and *Neither Close nor Frequent* categories for all the events $E_j \in \xi$.

Table 4.2 shows the top five entities in the *event specific* dictionaries for each event E_j and a socio-political event dictionary for the set of events ξ under study. The entities in the *event specific* dictionaries, when present in a source makes it highly specific to that particular event and contributes in gaining information about it. These entities can also help in conducting micro-analysis of an event by identifying precise details about the event. On the other hand, the entities of the *event-class specific* dictionary provides shallow information about a specific event, and are useful in learning about a category of events, in this case, socio-political uprisings in the middle east. These entities can help in conducting macro-analysis of an event by classifying the event into a certain category (like socio-political, crisis, entertainment, economic, etc.) and point out the general characteristics of the event. However, the analysis requires a set of known events, which could be a perceiveable constraint. We plan to address this in the future work.

4.8 Conclusions And Future Work

In this paper, we highlighted the need for exploring the social media sources to study an event that are often buried in the Long Tail. We demonstrated that social media sources have the capability to provide very specific and novel information. However, the sheer volume of social media sources and the Long Tail characteristics (e.g., link sparsity, colloquial language, etc.) make it extremely challenging to identify the specific sources. Towards this direction, we developed a methodology that utilizes relevant entities as a

mechanism to identify specific sources in a mutual reinforcement framework. Further, in order to consider the dynamic relationship between the specific sources and close entities, an evolutionary mutual reinforcement model is developed. Experiments conducted on real-world datasets for the three social movements during the Arab Spring, viz., Egyptian Revolution, Libyan Revolution, and Tunisian Revolution demonstrate faster convergence and better accuracy of the evolutionary mutual reinforcement model over the conventional mutual reinforcement model. Furthermore, the evolutionary mutual reinforcement model outperformed one of the most-widely used search engines, i.e., Google Blog Search and IceRocket. It was observed that the search engines ranked the specific sources surprisingly low, thereby reducing the chances of their discovery. The poor hyperlink connectivity of these Long Tail social media sources was contemplated to be a big reason behind their low ranks in traditional search engines. By analyzing the close entities identified by our model we also showed the potential of the framework to be utilized for analyzing events. In future, we plan to explore the proposed framework in assessing credibility of the Long Tail social media sources, which are known to disseminate false reports for breaking events. We also plan to use the framework in order to identify users with sustained interest in different events.

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 3.7 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 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 [35], and a source for gauging public interest and opinions [36]. 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 [22, 37, 38].

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 #CPACPick-upLines #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 [39]. 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' [40].

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?*
- *Which are the most informative tweets sharing event-specific information?*

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

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.

5.3 Challenges in Mining Tweets

5.3.1 Information Overload affecting Consumption and Collection of Data

As already pointed out previously that Twitter produces 500 million posts per day, that is distributed across its 288 million active users. This huge amount of content being generated increases the chances of an user to experience an overload of information that is left for him to consume. Surveys show that two thirds of Twitter users have felt that they receive too many posts, and over half of Twitter users have felt the need for a tool to filter out the irrelevant posts [41]. [42] also attempted to study and quantify information overload experienced by users in Twitter.

Due to the large volumes of information that needs to be stored and processed by Twitter on a daily basis, it also becomes a challenging job to build a infrastructure that can handle queries and information seeking needs of its millions of users. Therefore, the API endpoints provided by Twitter for collecting live data, results in collection of a biased sample. The Twitter Streaming API provides only a random sample of 1% of the total public stream produced in real-time, also known as spritzer sample. For accessing larger samples one needs to subscribe to Firehose services that are often costly and are seldom used by the researchers in an academic setting.

5.3.2 Idiosyncratic Structure and Informal Language

Unlike news documents and blogs, tweets pose additional challenges in the tasks of summarization, information retrieval, topic detection, entity extraction and POS tagging, due to the colloquial language used and limitation of 140 characters forcing the users to express the content in unusual grammatical structures that are not common in other publishing websites. The state-of-the-art entity extraction techniques depend on local linguistic features, common in well-formed documents like capitalization and POS tags of previous words [43]. Such characteristics is uncommon in tweets due to extensive use of informal language. Limitation in length, extensive and unusual usage of abbreviations,

capitalization and uncommon grammar makes it extremely difficult for the text mining techniques to identify valuable and useful content [3]. Due to live conversations between the users during real-life events it is also difficult to extract event related information as the conversation tweets often lack context. 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].

5.3.3 Nepotistic relationships

5.3.4 Data Management Challenges

5.4 Objective and Contributions

The main objective of the work presented in this chapter 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* [44], 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 model 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 Event Related Tweet Content

TABLE 5.2: Details of data collected for analyzing event related tweet content.

Event Name and Query Hashtag	**No. of Tweets**	**Time Period**
Sochi Winter Games 2014 (#sochi2014) ($http://goo.gl/sG4Rqd$)	1958220	11th Feb,2014 to 3rd March, 2014
SXSW 2014 (#sxsw2014) ($http://goo.gl/b6Nd6X$)	1880557	8th March, 2014 to 16th March, 2014
CPAC 2014 (#cpac2014) ($http://goo.gl/9o1KUx$)	18104	7th March, 2014 to 16th March, 2014

Given the mechanisms of user interactions and content production in Twitter as explained in Section 1, we analysed 3.8 million (approx) English tweets produced during three real-life events. Details of the data related to the events, collected for conducting the analysis is presented in Table 5.2[2]. We provided a popular hashtag corresponding to each event to the Twitter streaming API[3] in order to collect the data over the indicated period of time. The text of the tweets were preprocessed and prepared for analysis (Refer Section 5.8.2 for details). One of the main intentions behind this analysis was to investigate the nature of content in informative and non-informative tweets and to understand if there is a difference between tweets rich in generic information and the ones with event-specific information.

TABLE 5.3: Tweet features for content informativeness.

Has Url, No. of words, No. of stopwords, No. of feeling words, No. of slang words, No. of hashtags, No. of user mentions, Tweet length (No. of characters), No. of unique characters, No. of special characters, Favorite count, Retweet count, Formality, Is tweet verified, No. of nouns, No. of adjectives, No. of verbs, No. of adverbs, No. of pronouns, No. of interjections, No. of articles, No. of prepositions.

TABLE 5.4: Evaluation measures for logistic regression model.

	Precision	**Recall**	**F1-score**
Non-informative (0)	0.70	0.49	0.57
Informative (1)	0.78	0.90	0.84
Avg/Total	0.76	0.77	0.75
Accuracy = 76.64%			

[2] Note: This dataset is different from the dataset that we use for our experiment and evaluation. Refer Section 6.1 for the reason.

[3] https://dev.twitter.com/streaming/overview

FIGURE 5.1: Content characteristics of informative and non-informative tweets related to events.

		Average No. of Tokens	Average No. of Slang Words	Average Length	Average No. of Top Hashtags	Average No. of Top Nouns	Percentage of URLs
Sochi Winter Games 2014	*Informative*	8.55	0.47	115.55	0.44	5.14	96.32%
	Non-informative	3.55	0.77	69.92	1.23	1.78	1.04%
SXSW 2014	*Informative*	7.24	0.62	114.01	0.81	4.36	92.21%
	Non-informative	3.08	0.91	62.64	0.94	1.52	0.34%
CPAC 2014	*Informative*	6.81	0.53	126.83	1.84	2.42	76.01%
	Non-informative	3.55	0.9	88.65	2.04	2.04	0.68%

5.5.1 Analysis of Informative and Non-informative Content in Tweets

Our first step was to segregate the tweets likely to have informative content from the non-informative ones. In order to do so we trained a logistic regression model on an annotated dataset [45], which is publicly available. 9729 English language annotated tweets were used for building the model. The tweets labeled as *related and informative* were assigned a score of 1 and all the other tweets labeled as *related - but not informative* and *not related* were assigned a score of 0. Table 5.3 lists the features selected for each tweet. The choice of the features was driven by previous studies pointed out in Section 2. 10-fold cross validation was performed resulting in a model with an accuracy of 76.64%. Refer Table 5.4 for evaluation measures.

$$Formality = (No.\,of\,nouns + No.\,of\,adjectives + No.\,of\,prepositions + No.\,of\,articles - No.\,of\,pronouns - No.\,of\,verbs - No.\,of\,adverbs - No.\,of\,interjections + 100)/2$$

The logistic regression model was then used for assigning informativeness score between 0 and 1 to all the tweets in the dataset, with 0 being least informative and 1 being most informative. Although, the model is developed on tweets related to disaster events, it has been shown by the authors [46] that the annotations could be generalized to any type of event. The tweets for each event were then separated into two subsets - *informative*, containing tweets scoring more than 0.7, and *non-informative*, containing tweets scoring less than 0.3. Average values of different content characteristics of tweets were calculated for both the subsets. The top 10% of the frequently occurring hashtags, nouns and URLs were considered as top hashtags, nouns and URLs for the analysis, respectively. Some

of the characteristics that were noticeably different for informative and non-informative tweets are listed in Table **??**.

For all the events, on an average, the informative tweets were marked by a higher number of words per tweet and greater occurrence of top nouns. The average length of informative tweets were also more than the non-informative ones. The percentage of informative tweets having URLs were strikingly high. As expected, a greater usage of slang words was observed in non-informative tweets. However, the greater occurrence of top hashtags in non-informative tweets urged us to look into the content. We found that a lot of non-informative tweets have used many popular hashtags with unrelated content and URLs pointing to irrelevant information. This is typical of spam tweets as already reported by [47]. Not shown due to space constraints, a larger average occurrence of feeling words and top URLs was observed in informative tweets. The average number of follower counts for users posting informative tweets was also observed to be higher than the ones posting non-informative ones.

The above observations gave us an idea of how informative content about events is generally produced in Twitter and the characteristics that differentiates it from non-informative ones. It is now intuitive that the informative tweets are more expressive, formal and lengthier, marked by higher presence of nouns. Due to the constraints imposed by Twitter on the number of characters in a tweet, the users tend to share URLs along with the textual content that might lead to more information about the event. Also, users with high follower counts tend to post informative tweets. This is intuitive, as the tweets posted by such users are read by a larger audience. This might encourage them to share informative content. Also, it might be that since they share informative content, they are followed by large number of users.

5.5.2 Difference between Informative and Event-specific Informative Tweets:

Our second step was to manually analyze the informative tweets and understand if it is good enough to train a classifier for detecting informative tweets for an event in order to identify valuable event-specific information. Although the tweets on which we trained our logistic regression model were related to events yet we came across tweets like, *RT @BFDealz: http://t.co/TSJAigrVJI WHEELS SUPER TREASURE HUNT SUPERIZED HARLEY DAVIDSON FAT BOY LONG CARD 2014 #cpac2014 #sxsw*, which were classified as informative, even when it did not contain any event-specific information.

This was probably because of the choice of features for the model, which were generic and not event-specific. The model did not take into account the presence of features that were popular and specific to the events, like popular hashtags, text units, etc. Popularity alone might not work as it is often mis-used by the spammers. It is also challenging to come up with a list of such event-specific features. Moreover, if one can compile such a list then it would be difficult to set thresholds on each such feature in order to qualify it as event-specific. Also, a supervised classification model does not have the ability to simultaneously rank tweets, hashtags, text units, URLs and users in terms of event-specific informativeness. After going through the existing literature we assume that the challenges discussed above would be a shortcoming of any supervised model and there is a need for an alternative feasible approach. It is also difficult to predict the event-specific informativeness in the URLs shared along with the tweets, as it might be necessary to analyze the content pointed to by the URLs. Also, not all the URLs contain text. They might be images or videos providing valuable information about an event. This motivated us to devise a novel framework that solves all the above problems and is discussed in Section 5.

5.6 Problem Statement

In this section, we give the definition of an event appropriate in the context of our problem, and then present a formal statement of the problem that we want to solve.

Events have been defined from various perspectives and in different contexts. In the context of our work we adopt a definition similar to [48].

Event: An event is defined as a real-world occurrence (E_i) with an associated time period $_{T_{E_i}}$ ($t_{E_i}^{start}$-$t_{E_i}^{end}$) and a time ordered stream of tweets M_{E_i}, of substantial volume, discussing about the event and posted in time T_{E_i}. The tweets are primarily composed of a set of hashtags (H_{E_i}) used for annotating the tweets ($\in M_{E_i}$), a set of text units (W_{E_i}) used for sharing textual information in the tweets ($\in M_{E_i}$), a set of URLs (L_{E_i}) linking to external sources related to the event and a set of users (U_{E_i}) posting the tweets ($\in M_{E_i}$).

Problem: *Given an event E_i, a time ordered stream of n tweets $M_{E_i} = \{m_1, ..., m_i, m_j, ..., m_n\}$ related to the event posted in time period T_{E_i}, a set of hashtags $H_{E_i} = \{h_1, h_2, ..., h_p\}$, a set of text units $W_{E_i} = \{w_1, w_2, ..., w_r\}$, a set of URLs $L_{E_i} = \{l_1, l_2, ..., l_t\}$ and a set of users $U_{E_i} = \{u_1, u_2, ..., u_s\}$, the problem is to find a ranked set of:*

- *tweets* $\hat{M}_{E_i} = \{m_1 \geq ... \geq m_i \geq m_j \geq ... \geq m_n \mid i < j\}$,

- *hashtags* $\hat{H}_{E_i} = \{h_1 \geq ... \geq h_i \geq h_j \geq ... \geq h_p \mid i < j\}$,
- *text units* $\hat{W}_{E_i} = \{w_1 \geq ... \geq w_i \geq w_j \geq ... \geq w_r \mid i < j\}$,
- *URLs* $\hat{L}_{E_i} = \{l_1 \geq ... \geq l_i \geq l_j \geq ... \geq l_t \mid i < j\}$,
- *users* $\hat{U}_{E_i} = \{u_1 \geq ... \geq u_i \geq u_j \geq ... \geq u_s \mid i < j\}$,

ordered in decreasing order of its event-specific informativeness.

5.7 Methodology

In this section, we present *TwitterEventInfoGraph*, a graph structure representing implicit mutually reinforcing relationships between event related *tweets*, *hashtags*, *text units*, *URLs*, and *users*. We explain and quantify the semantics of the relationships between the nodes of the graph. Then we devise an algorithm - *TwitterEventInfoRank*, for ranking the nodes of the graph leveraging the mutually reinforcing relationships between them. All this constitute a novel framework for simultaneous identification and ranking of *tweets*, *hashtags*, *text units*, *URLs*, and *users* in terms of event-specific informativeness, which we present next.

TABLE 5.5: Affinity scores of edges between vertices of TwitterEventInfoGraph

Affinity scores (edge weights) between different vertices $\in M_{E_i}, H_{E_i}, W_{E_i}, U_{E_i}, L_{E_i}$***:***

$P(h_i \mid w_j) = \frac{No.\,of\,tweets\,h_i\,and\,w_j\,occur\,together}{No.\,of\,tweets\,w_j\,occurs}$, $P(w_i \mid h_j) = \frac{No.\,of\,tweets\,w_i\,and\,h_j\,occur\,together}{No.\,of\,tweets\,h_j\,occurs}$,
$P(h_i \mid l_j) = \frac{No.\,of\,tweets\,h_i\,and\,l_j\,occur\,together}{No.\,of\,tweets\,l_j\,occurs}$, $P(l_i \mid h_j) = \frac{No.\,of\,tweets\,l_i\,and\,h_j\,occur\,together}{No.\,of\,tweets\,h_j\,occurs}$,
$P(h_i \mid u_j) = \frac{No.\,of\,tweets\,h_i\,and\,u_j\,occur\,together}{No.\,of\,tweets\,u_j\,occurs}$, $P(u_i \mid h_j) = \frac{No.\,of\,tweets\,u_i\,and\,h_j\,occur\,together}{No.\,of\,tweets\,h_j\,occurs}$,
$P(w_i \mid l_j) = \frac{No.\,of\,tweets\,w_i\,and\,l_j\,occur\,together}{No.\,of\,tweets\,l_j\,occurs}$, $P(l_i \mid w_j) = \frac{No.\,of\,tweets\,l_i\,and\,w_j\,occur\,together}{No.\,of\,tweets\,w_j\,occurs}$,
$P(w_i \mid u_j) = \frac{No.\,of\,tweets\,w_i\,and\,u_j\,occur\,together}{No.\,of\,tweets\,u_j\,occurs}$, $P(u_i \mid w_j) = \frac{No.\,of\,tweets\,u_i\,and\,w_j\,occur\,together}{No.\,of\,tweets\,w_j\,occurs}$,
$P(u_i \mid l_j) = \frac{No.\,of\,tweets\,u_i\,and\,l_j\,occur\,together}{No.\,of\,tweets\,l_j\,occurs}$, $P(l_i \mid u_j) = \frac{No.\,of\,tweets\,l_i\,and\,u_j\,occur\,together}{No.\,of\,tweets\,u_j\,occurs}$,
$P(h_i \mid m_j) = P(m_i \mid h_j) = P(w_i \mid m_j) = P(m_i \mid w_j) = P(u_i \mid m_j) = P(m_i \mid u_j) = P(l_i \mid m_j) = P(m_i \mid l_j) = 1.0$

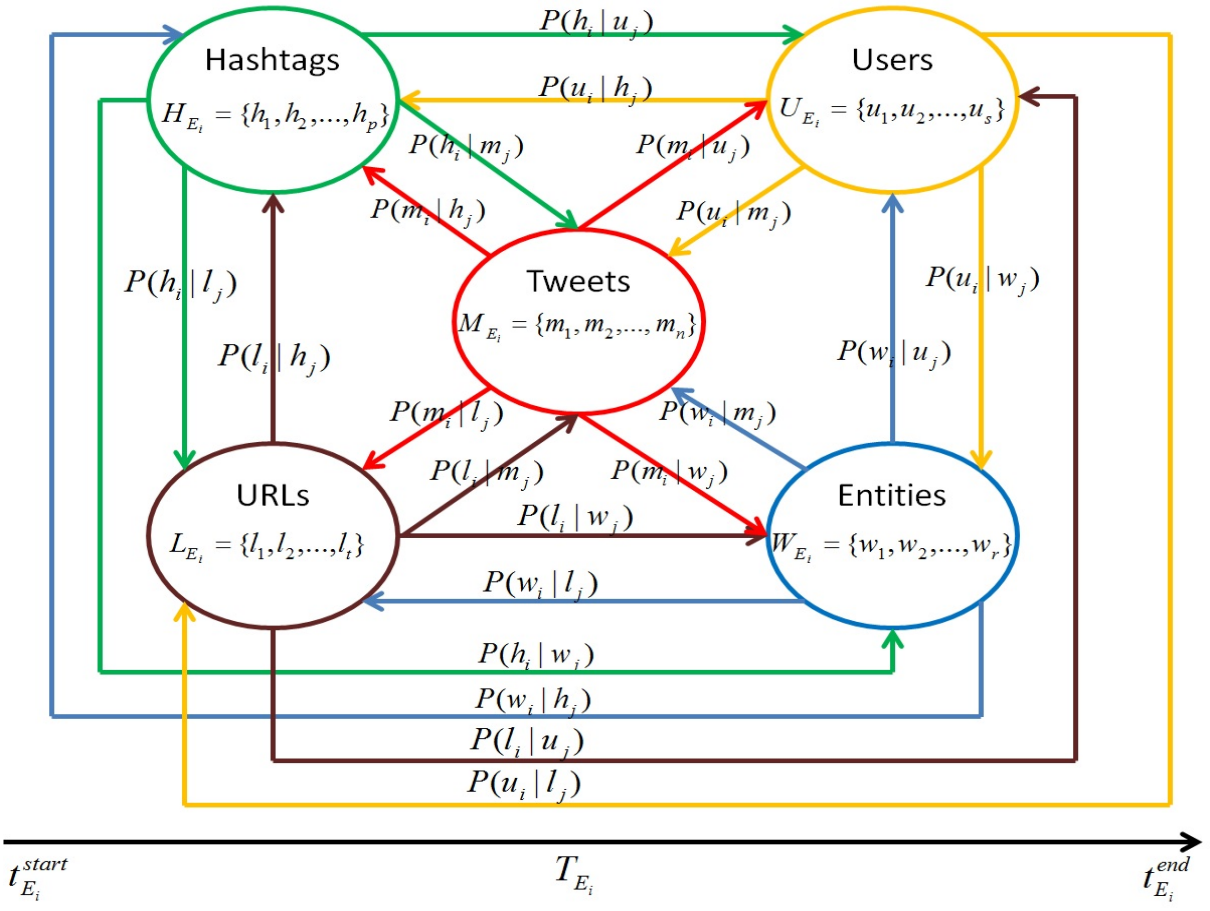
Note: $P(h_i \mid w_j)$ should be read as the probability of occurrence of hashtag h_i given the occurrence of the text unit w_j in the stream of tweets M_{E_i} related to event E_i collected over the time period T_{E_i}. Similarly, for others.

5.7.1 TwitterEventInfoGraph

After the observations in the previous section we conclude that the informative tweets in general are characterized by wordiness, occurrences of URLs and are posted by users with

high follower count. These characteristics are also the primary features that distinguish informative from non-informative content. Although, presence of hashtags is not a good indicator of informativeness, yet it is a strong identifier of a topic as already pointed by [39]. Popular hashtags for an event might be used maliciously. On the other hand, the presence of a popular hashtag in a wordy tweet consisting of words popular for the event, along with a popular URL, posted by an influential user is highly likely to contain event-specific content. Therefore, it is intuitive that given a stream of tweets for an event an optimal combination of event related popular text units (words, unigrams, bigrams etc), hashtags, and URLs, posted by an influential user in a tweet, is one of the key indicators for identifying event-specific informative content. It would be highly unlikely for a tweet to contain all of these and yet not convey useful event-specific information. Building on this intuition we model our framework based on the following assumptions:

FIGURE 5.2: Mutual Reinforcement Chains in Twitter for an event.



For an event E_i

- a *tweet is an event-specific informative tweet* if it is strongly associated with:

  (a) *event-specific informative hashtags*,

(b) *event-specific informative text units*,

(c) *event-specific informative users*,

(d) *event-specific informative URLs*.

- a *hashtag is an event-specific informative hashtag* if it is strongly associated with:

  (a) *event-specific informative tweets*,

  (b) *event-specific informative text units*,

  (c) *event-specific informative users*,

  (d) *event-specific informative URLs*.

- a *text unit is an event-specific informative text unit* if it is strongly associated with:

  (a) *event-specific informative tweets*,

  (b) *event-specific informative hashtags*,

  (c) *event-specific informative users*,

  (d) *event-specific informative URLs*.

- a *user is an event-specific informative user* if it is strongly associated with:

  (a) *event-specific informative tweets*,

  (b) *event-specific informative hashtags*,

  (c) *event-specific informative text units*,

  (d) *event-specific informative URLs*.

- a *URL is an event-specific informative URL* if it is strongly associated with:

  (a) *event-specific informative tweets*,

  (b) *event-specific informative hashtags*,

  (c) *event-specific informative text units*,

  (d) *event-specific informative users*.

The relationships for an event E_i as stated above, forms a *Mutual Reinforcement Chain* [44] for the event E_i as shown in Figure 5.2. We represent this relationship in a graph $\mathbf{G = (V,D)}$, which we call as *TwitterEventInfoGraph*, where $\mathbf{V} = \mathbf{M_{E_i}} \cup \mathbf{H_{E_i}} \cup \mathbf{W_{E_i}} \cup \mathbf{U_{E_i}} \cup \mathbf{L_{E_i}}$, is the set of vertices and $\mathbf{D}$ is the set of directed edges between different vertices.

Whenever two vertices are associated, there are two edges between them that are oppositely directed. Each directed edge is assigned a weight, which determines the degree of association of one vertex with the other. The weights for each edge is calculated according to the conditional probabilities given in Table 5.5.

We do not consider an edge between two vertices of same type. That is, we don't connect a tweet with another tweet. Similarly, for hashtags, text units, users and URLs. This constraint was imposed in order to deal with the nepotistic relationships between high quality content and low quality content introduced by the malicious users for promoting the low quality content. We observe these malicious side effects in the results obtained for *TextRank* explained in Section 6.5.

Next, we explain *TwitterEventInfoRank*.

5.7.2 TwitterEventInfoRank

In this section, we introduce an iterative algorithm that takes into account the mutually reinforcing relationships between the vertices of *TwitterEventInfoGraph* as explained in the previous section and propagates event-specific scores of each vertex to connected vertices across the graph for ranking its vertices ($\in V$) in terms of event-specific informativeness.

We first assign a event-specific score to all the vertices of the graph. Event-specific scores for vertices ($\in H_{E_i}, W_{E_i}, U_{E_i}, L_{E_i}$) are calculated using equations (1-4) as presented in Table 5.5. The tweets ($\in M_{E_i}$) are assigned an initial informativeness score as obtained from the logistic regression model explained in Section 3. The event-specific scores for vertices ($\in H_{E_i}, W_{E_i}, U_{E_i}, L_{E_i}$) and informativeness score for vertices ($\in M_{E_i}$) gives an initial ranking of all the vertices of *TwitterEventInfoGraph*. We aim to refine the initial scores and assign a final score for ranking the vertices by leveraging the mutually reinforcing relationships between them.

$$Score(h_i) = \frac{freq(h_i)}{max\{freq(h_1), freq(h_2), ..., freq(h_p)\}} \tag{5.1}$$

$$Score(w_i) = \frac{freq(w_i)}{max\{freq(w_1), freq(w_2), ..., freq(w_r)\}} \tag{5.2}$$

$$Score(u_i) = \frac{followers(u_i)}{max\{followers(u_1), ..., followers(u_r)\}} \tag{5.3}$$

$$Score(l_i) = \frac{freq(l_i)}{max\{freq(l_1), freq(l_2), ..., freq(l_r)\}} \tag{5.4}$$

The relationships between two different subsets of vertices in graph $\mathbf{G}$ is denoted by an affinity matrix. For e.g., $\mathbf{A^{MH}_{E_i}}$ denotes the $\mathbf{M_{E_i} - H_{E_i}}$ affinity matrix for event E_i, where $(\mathbf{i,j})^{\mathbf{th}}$ entry is the edge weight quantifying the association between i^{th} tweet ($\in M_{E_i}$) and j^{th} hashtag ($\in H_{E_i}$), calculated using Table 5.5. Similarly, $\mathbf{A^{WH}_{E_i}}$ denotes the $\mathbf{W_{E_i} - H_{E_i}}$ affinity matrix between set of text units W_{E_i} and set of hashtags H_{E_i} for event E_i, and so on.

The rankings of *tweets*, *hashtags*, *text units*, *users* and *URLs* in terms of event-specific informativeness, can be iteratively derived from the Mutual Reinforcement Chain for the event. Let $R^M_{E_i}$, $R^H_{E_i}$, $R^W_{E_i}$, $R^U_{E_i}$ and $R^L_{E_i}$ denote the ranking scores for the set of tweets ($\in M_E$), set of hashtags ($\in H_{E_i}$), set of text units ($\in W_{E_i}$), set of users ($\in U_{E_i}$), and set of URLs ($\in L_{E_i}$), respectively. Therefore, the Mutual Reinforcement Chain ranking for the k^{th} iteration can be formulated as follows:

$$R^{M(k+1)}_{E_i} = A^{MM(k)}_{E_i} R^{M(k)}_{E_i} + A^{MH(k)}_{E_i} R^{H(k)}_{E_i} + A^{MW(k)}_{E_i} R^{W(k)}_{E_i} + A^{MU(k)}_{E_i} R^{U(k)}_{E_i} + A^{ML(k)}_{E_i} R^{L(k)}_{E_i} \tag{5.5}$$

$$R^{H(k+1)}_{E_i} = A^{HM(k)}_{E_i} R^{M(k)}_{E_i} + A^{HH(k)}_{E_i} R^{H(k)}_{E_i} + A^{HW(k)}_{E_i} R^{W(k)}_{E_i} + A^{HU(k)}_{E_i} R^{U(k)}_{E_i} + A^{HL(k)}_{E_i} R^{L(k)}_{E_i} \tag{5.6}$$

$$R^{W(k+1)}_{E_i} = A^{WM(k)}_{E_i} R^{M(k)}_{E_i} + A^{WH(k)}_{E_i} R^{H(k)}_{E_i} + A^{WW(k)}_{E_i} R^{W(k)}_{E_i} + A^{WU(k)}_{E_i} R^{U(k)}_{E_i} + A^{WL(k)}_{E_i} R^{L(k)}_{E_i} \tag{5.7}$$

$$R^{U(k+1)}_{E_i} = A^{UM(k)}_{E_i} R^{M(k)}_{E_i} + A^{UH(k)}_{E_i} R^{H(k)}_{E_i} + A^{UW(k)}_{E_i} R^{W(k)}_{E_i} + A^{UU(k)}_{E_i} R^{U(k)}_{E_i} + A^{UL(k)}_{E_i} R^{L(k)}_{E_i} \tag{5.8}$$

$$R^{L(k+1)}_{E_i} = A^{LM(k)}_{E_i} R^{M(k)}_{E_i} + A^{LH(k)}_{E_i} R^{H(k)}_{E_i} + A^{LW(k)}_{E_i} R^{W(k)}_{E_i} + A^{LU(k)}_{E_i} R^{U(k)}_{E_i} + A^{LL(k)}_{E_i} R^{L(k)}_{E_i} \tag{5.9}$$

The equations 5-9 can be represented in the form of a block matrix Δ_{E_i}, where,

$$\Delta_{E_i} = \begin{pmatrix} A_{E_i}^{MM} & A_{E_i}^{MH} & A_{E_i}^{MW} & A_{E_i}^{MU} & A_{E_i}^{ML} \\ A_{E_i}^{HM} & A_{E_i}^{HH} & A_{E_i}^{HW} & A_{E_i}^{HU} & A_{E_i}^{HL} \\ A_{E_i}^{WM} & A_{E_i}^{WH} & A_{E_i}^{WW} & A_{E_i}^{WU} & A_{E_i}^{WL} \\ A_{E_i}^{UM} & A_{E_i}^{UH} & A_{E_i}^{UW} & A_{E_i}^{UU} & A_{E_i}^{UL} \\ A_{E_i}^{LM} & A_{E_i}^{LH} & A_{E_i}^{LW} & A_{E_i}^{LU} & A_{E_i}^{LL} \end{pmatrix}$$

Let

$$R_{E_i} = \begin{pmatrix} R_{E_i}^{M} \\ R_{E_i}^{H} \\ R_{E_i}^{W} \\ R_{E_i}^{U} \\ R_{E_i}^{L} \end{pmatrix}$$

then, R_{E_i} can be computed as the dominant eigenvector of Δ_{E_i}.

$$\Delta_{E_i}.R_{E_i} = \lambda.R_{E_i} \tag{5.10}$$

In order to guarantee a unique R_{E_i}, Δ_{E_i} must be forced to be stochastic and irreducible.

To make Δ_{E_i} stochastic we divide the value of each element in a column of Δ_{E_i} by the sum of the values of all the elements in that column. This finally makes Δ_{E_i} column stochastic. We now denote it by $\hat{\Delta}_{E_i}$.

Next, we make $\hat{\Delta}_{E_i}$ irreducible. This is done by making the graph G strongly connected by adding links from one node to any other node with a probability vector p. Now, $\hat{\Delta}_{E_i}$ is transformed to

$$\overline{\Delta}_{E_i} = \alpha\hat{\Delta}_{E_i} + (1-\alpha)E \tag{5.11}$$

$$E = p \times [1]_{1\times k} \tag{5.12}$$

where $0 \leq \alpha \leq 1$ is set to 0.85 according to *PageRank*, and k is the order of $\hat{\Delta}_{E_i}$. We set $p = [1/k]_{k\times 1}$ by assuming a uniform distribution over all elements. Now, $\overline{\Delta}_{E_i}$ is stochastic and irreducible and it can be shown that it is also primitive by checking $\overline{\Delta}_{E_i}^2$ is greater than 0.

Following steps are taken next,

1. We initialize the rank vectors $(R_{E_i}^{M(0)}, R_{E_i}^{H(0)}, R_{E_i}^{W(0)}, R_{E_i}^{U(0)}, R_{E_i}^{L(0)})$ for each subset of vertices $(M_{E_i}, H_{E_i}, W_{E_i}, U_{E_i}, L_{E_i})$. We use the event-specific scores calculated

for the set of hashtags, text units, users and urls as their initial scores. All the scores lie between 0 and 1. For the tweets we use the logistic regression model and assign each one of them an initial informativeness score between 0 and 1.

2. Then we assign

$$R^0_{E_i} = \begin{pmatrix} R^{M(0)}_{E_i} \\ R^{H(0)}_{E_i} \\ R^{W(0)}_{E_i} \\ R^{U(0)}_{E_i} \\ R^{L(0)}_{E_i} \end{pmatrix}$$

and normalize $R^0_{E_i}$ such that $|| R^0_{E_i} ||_1 = 1$

3. Apply power iteration method using the same parameters as used in PageRank with the convergence tolerance set at 1e-08 and $\lambda = 0.85$.

4. We get the final rank vectors for each subset of the vertices $(R^M_{E_i}, R^H_{E_i}, R^W_{E_i}, R^U_{E_i}, R^L_{E_i})$ after convergence.

5. We finally obtain the subsets $\hat{M}_{E_i}, \hat{H}_{E_i}, \hat{W}_{E_i}, \hat{L}_{E_i}, \hat{U}_{E_i}$ consisting of the *tweets*, *hashtags*, *text units*, *URLs* and *users*, respectively arranged in descending order of their final scores.

The final ordered subsets $\hat{\mathbf{M}}_{\mathbf{E_i}}$,$\hat{\mathbf{H}}_{\mathbf{E_i}}$,$\hat{\mathbf{W}}_{\mathbf{E_i}}$,$\hat{\mathbf{L}}_{\mathbf{E_i}}$,$\hat{\mathbf{U}}_{\mathbf{E_i}}$, thus obtained are the tweets, hashtags, text units, URLs and users, ranked in terms of their event-specific informativeness.

During the implementation of the *TwitterEventInfoRank* algorithm the slang hashtags were removed. We only considered nouns as the text units and removed the slang words. We already reported in our analysis that non-informative tweets have higher slang content. Therefore, removal of slang hashtags and text units was done in order to obtain high quality results. We also showed higher occurrence of nouns in informative tweets. Also, the occurrence of a noun in a tweet intuitively suggests that the tweet has information about a person, place, or thing. Thus, we only considered the set of nouns extracted from the tweets as the set of text units.

The text units are generic units in the framework and can be changed according to specific requirements. Entities extracted from the textual content of tweets could be experimented, in place of nouns. Since the algorithm uses power iteration method for ranking the vertices of the graph, it could be easily made scalable using mapreduce paradigm [49]. We plan to work on it in the future and implement our framework using hadoop and mapreduce environment.

Input : Sets of vertices $M_{E_i}, H_{E_i}, W_{E_i}, U_{E_i}, L_{E_i}$ of graph G, $\alpha = 0.85$, $\varepsilon = 1e-08$.

Output: Ordered set of vertices $\hat{M}_{E_i}$, containing tweets ranked in order of event-specific informative content sharing information about event related entities.

Steps:

Initialize rank vectors $[R_{E_i}^{M(0)}, R_{E_i}^{H(0)}, R_{E_i}^{W(0)}, R_{E_i}^{U(0)}, R_{E_i}^{L(0)}]$;

Assign $R_{E_i}^0 = [R_{E_i}^{M(0)}, R_{E_i}^{H(0)}, R_{E_i}^{W(0)}, R_{E_i}^{U(0)}, R_{E_i}^{L(0)}]^T$;

Normalize $R_{E_i}^0$ such that $|| R_{E_i}^0 ||_1 = 1$;

Construct matrix Δ_{E_i};

Make matrix Δ_{E_i} stochastic and irreducible converting it to $\overline{\Delta}_{E_i}$;

$k \leftarrow 1$

repeat

  $R_{E_i}^k \leftarrow \overline{\Delta}_{E_i} R_{E_i}^{k-1}$;

  $k \leftarrow k + 1$;

until $|| R_{E_i}^k - R_{E_i}^{k-1} ||_1 < \varepsilon \; OR \; k \geq 100$;

$R_{E_i}^M \leftarrow R_{E_i}^{M(k)}$, $R_{E_i}^H \leftarrow R_{E_i}^{H(k)}$, $R_{E_i}^W \leftarrow R_{E_i}^{W(k)}$, $R_{E_i}^U \leftarrow R_{E_i}^{U(k)}$, $R_{E_i}^L \leftarrow R_{E_i}^{L(k)}$;

$\hat{M}_{E_i} \leftarrow R_{E_i}^M$, $\hat{H}_{E_i} \leftarrow R_{E_i}^H$, $\hat{W}_{E_i} \leftarrow R_{E_i}^W$, $\hat{U}_{E_i} \leftarrow R_{E_i}^U$, $\hat{L}_{E_i} \leftarrow R_{E_i}^L$;

return $\hat{M}_{E_i}$,$\hat{H}_{E_i}$,$\hat{W}_{E_i}$,$\hat{U}_{E_i}$,$\hat{L}_{E_i}$;

Since, our proposed framework takes a hybrid approach by using both supervised and unsupervised component, it is easily applicable in situations where an event needs to be tracked over time. The supervised portion assigns an initial generic informativeness score to the tweets for bootstrapping an unsupervised process that finally assigns event-specific informativeness scores. When applied over a time period the method for assigning the initial supervised scores might remain the same and the unsupervised process can change the rankings of the tweet contents as the event evolves.

Next, we present details of experimental settings and evaluation.

5.8 Experimental Settings and Evaluation

In this section, we explain the settings of the experiment conducted by us. We give the details of the data collected for performing the experiment. All the data preparation steps used for preprocessing the tweets before analyzing them and applying the algorithms are explained. We present the baselines used for comparing the effectiveness of our proposed algorithm and perform the evaluation tasks. We go through the evaluation results and discuss about the performance of our algorithm.

TABLE 5.6: Details of data collected for the experiment.

Event Name and Query Hashtag	**No. of Tweets**	**Time Period (UTC)**
Millions March NYC (#millionsmarchnyc) ($http://goo.gl/I8WR4B$)	56927	13th Dec, 2014 20:25:43 to 14th Dec, 2014 03:30:41
Sydney Siege (#sydneysiege) ($http://goo.gl/qLguvG$)	398204	15th Dec, 2014 07:21:16 to 15th Dec, 2014 22:46:45

Next, we present details of the data collected for the experiment.

5.8.1 Data Collection

For implementing and evaluating our proposed algorithm we collected 455,131 tweets from two real-life events, 'Millions March NYC' and 'Sydney Siege', using Twitter Streaming API. Details of the dataset is presented in Table 5.6. Tweets for each event was collected over the given period of time, by providing a popular hashtag corresponding to each event to the Twitter streaming API. The events for the experiments are different from the events selected for initial analysis as the choice was driven by its availability in Seen.co event database, whose ranking scores[4] are used as one of the baselines representing the state-of-the-art technique.

5.8.2 Data Preparation

We performed a series of data preparation steps before analyzing the tweets and implementing the *TwitterEventInfoRank* algorithm. Tweets having duplicate content were detected using md5 hashing scheme, and redundant copies were filtered out keeping a single representation of the tweet in our database. Although, the methodology is language independent, we only considered English language tweets, as the manual annotators used for evaluation were only proficient in English. Also the natural language toolkits used for the work gave best results for English text.

We used the default parts-of-speech (POS) tagging module provided by NLTK library[5]. A standard list of english stop words was used for eliminating the stop words from tweet

[4] Tweets in Seen.co is ranked according to their proprietary algorithm SeenRank and the scores are available in the response of their API found at (http://developer.seen.co/) We used a python wrapper freely available at https://github.com/dxmahata/pySeen for collecting data from Seen.co

[5] http://nltk.org

text. All the characters of the tweets were converted to lower case. The tweets were tokenized after detecting the POS tags and removing the special characters. We filtered out the user mentions, retweet symbol (*RT*) and URLs from the text during tokenization and did not consider them as tokens. A list of words expressing feelings was obtained from *wefeelfine.org*. Twitter related slang words were obtained from a publicly available document published by United States FBI[6]. A final list of slang words was compiled by adding some more internet slangs. The list would be made available on request. Retweet counts, favorite counts, verification information, user followers count and time information were obtained from the metadata attached with each tweet returned by Twitter API. The URLs shared in tweets are generally shortened. Due to the use of different URL shortener services, a single URL might be represented in different forms by each service. In order to solve this problem, we used AlchemyAPI[7] to expand the URLs to their original form.

5.8.3 Experiment with Named Entities as Text Units

5.8.3.1 Baselines

5.8.4 Experiment with Nouns as Text Units

5.8.4.1 Baselines

In order to evaluate the performance of *TwitterEventInfoRank* we selected six different algorithms that acted as our baselines. Please refer to the *Related Work* section (Section 2) for pointers to scientific literature explaining the techniques. Three of the baseline algorithms *LexRank*, *TextRank*, *Centroid* are widely used by the scientific community. Among the other three, one of them is a proprietary algorithm known as *SeenRank* commercially used by Seen.co for generating event summaries and highlights from Twitter. We considered *SeenRank* as the state-of-the-art technique. The other one is the Logistic Regression model that we implemented for initializing the informativeness score of the tweets. We considered it in order to make sure that our algorithm improves upon the initial generic informativeness score already assigned to the tweets at the start of the iteration and assigns event-specific informativeness scores on convergence. Number of retweets is a good measure of popularity of a tweet and is also used by Twitter for ranking its search results. Therefore, we also considered tweets ordered in decreasing order of number of retweets as one of our baselines. We name this scheme as *RTRank*

[6] https://www.documentcloud.org/documents/1199460-responsive-documents.html#document/p1
[7] http://alchemyapi.com

Centroid is one of the techniques that was previously used in the literature for solving a part of our problem that ranks tweets. In order to implement it as a baseline we considered the tweets for the event in the given time period as one cluster. After pre-processing the tweets, we calculated the centroid of the cluster and ordered the tweets in the decreasing order of their similarities with the centroid. We selected *LexRank* and *TextRank*, as they are graph-based techniques for ranking textual documents, and are widely used for generating summaries. We used the open-source implementation of *LexRank* available with sumy[8] package. In case of *TextRank*, we modified the algorithm in order to make it suitable for our context. Apart from creating heterogeneous relationships in *TwitterEventInfoGraph* we also created homogeneous relationships between the *information units* as well as the tweets. Cosine similarity (≥ 0.10) was used as the measure of relatedness between tweets, and the association scores of the hashtags, text units, users and URLs were based on their co-occurrence normalized between 0 and 1. The users were associated whenever they mentioned each other in the tweets, and the association score was measured by the number of mentions normalized between 0 and 1.

Due to unavailability of proper baseline techniques for ranking hashtags, text units, URLs and users in terms of event-specific informativeness we do not compare the results obtained for them with any other approach. However, we report their average scores and sample results.

TABLE 5.7: Avg IIC scores and total avg scores of annotations for Millions March NYC event.

Millions March NYC	**IIC**	**Total Avg Score (1-3)**
Top 50 event-specific informative Hashtags	0.786	1.980
Top 50 event-specific informative Text Units	0.880	1.320
Top 50 event-specific informative URLs	0.926	2.560
Top 50 event-specific informative Users	0.700	2.386
Top 100 event-specific informative Tweets	0.760	2.59

5.8.4.2 Evaluation

Evaluation Setup and Objectives: We evaluated the rankings obtained using *TwitterEventInfoRank* on the two datasets by comparing its performance with the selected baselines. A subset of tweets for each event for a given time period (one hour) was

[8] https://pypi.python.org/pypi/sumy/0.1.0

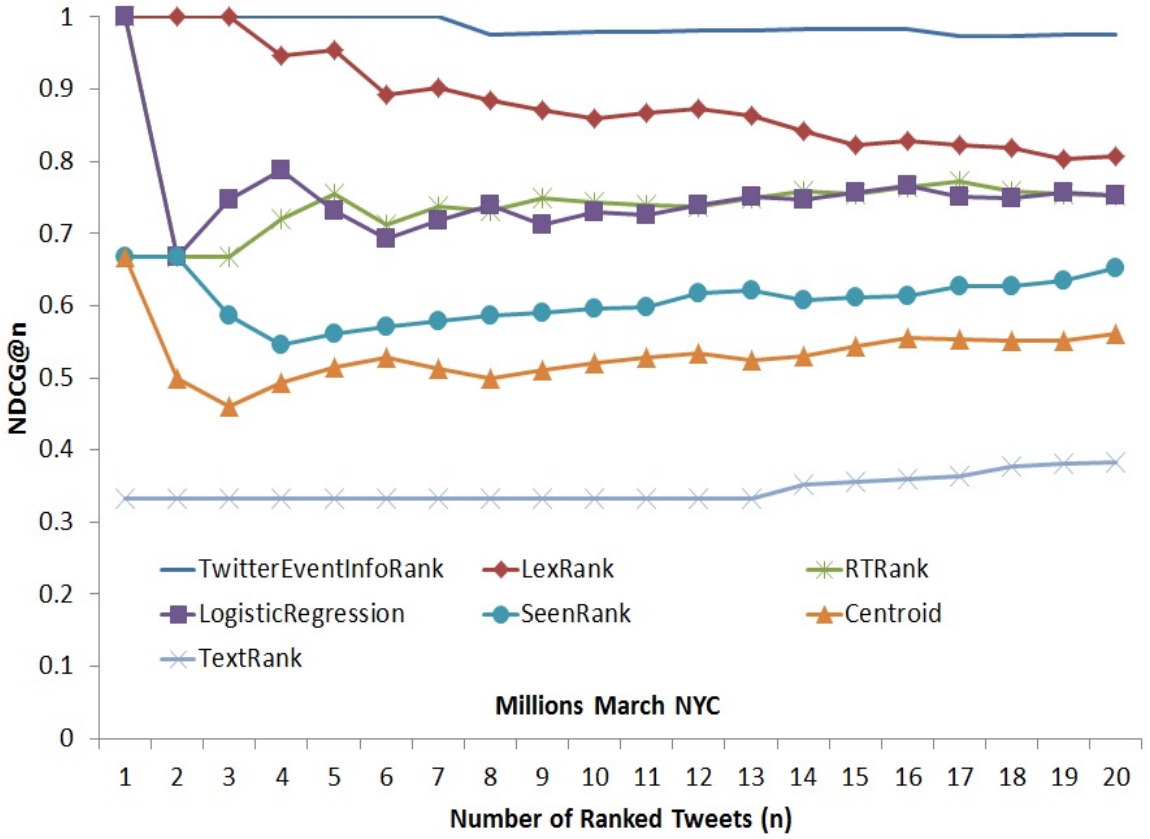


FIGURE 5.3: Performance comparison of ranking techniques using NDCG scores.

TABLE 5.8: Avg IIC scores and total avg scores of annotations for Sydney Siege event.

Sydney Siege	**IIC**	**Total Avg Score (1-3)**
Top 50 event-specific informative Hashtags	0.880	2.027
Top 50 event-specific informative Text Units	0.986	1.487
Top 50 event-specific informative URLs	0.893	2.413
Top 50 event-specific informative Users	0.646	2.353
Top 100 event-specific informative Tweets	0.83	2.62

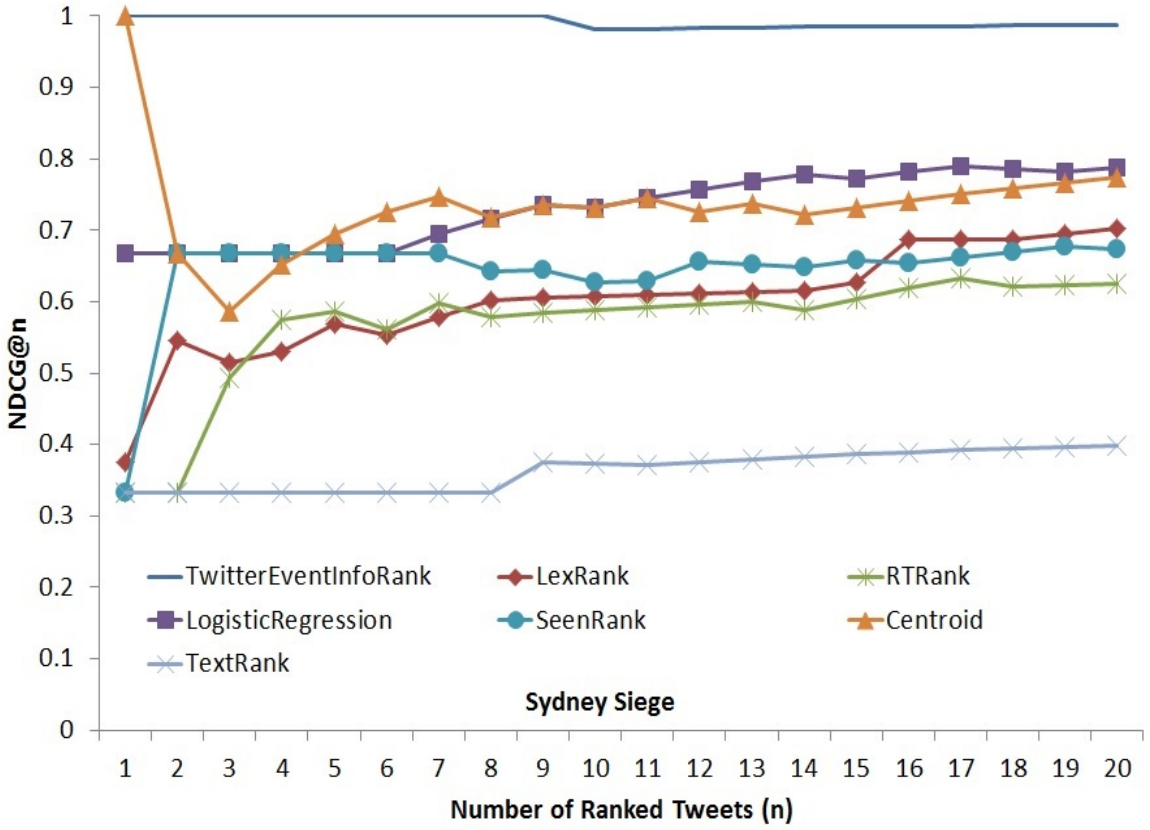


FIGURE 5.4: Performance comparison of ranking techniques using NDCG scores.

Technique	@ 10	@ 20	@ 30	@ 40	@ 50	@ 60	@ 70	@ 80	@ 90	@ 100
TwitterEventInfoRank	0.979	0.975	0.966	0.966	0.957	0.936	0.951	0.960	0.967	0.989
LexRank	0.859	0.807	0.830	0.813	0.822	0.825	0.834	0.878	0.922	0.944
RTRank	0.744	0.752	0.749	0.765	0.792	0.822	0.861	0.870	0.884	0.922
Logistic Regression	0.729	0.753	0.757	0.752	0.757	0.776	0.792	0.839	0.878	0.915
SeenRank	0.595	0.652	0.708	0.733	0.745	0.759	0.801	0.828	0.859	0.884
Centroid	0.519	0.560	0.623	0.658	0.690	0.727	0.747	0.788	0.835	0.857
TextRank	0.333	0.383	0.418	0.468	0.499	0.564	0.633	0.681	0.729	0.782

FIGURE 5.5: Performance comparison of ranking techniques using NDCG scores.

Technique	@ 10	@ 20	@ 30	@ 40	@ 50	@ 60	@ 70	@ 80	@ 90	@ 100
TwitterEventInfoRank	0.980	0.987	0.968	0.957	0.954	0.941	0.946	0.952	0.960	0.990
LexRank	0.607	0.701	0.684	0.707	0.737	0.768	0.764	0.806	0.838	0.868
RTRank	0.588	0.624	0.677	0.716	0.729	0.751	0.769	0.821	0.863	0.880
Logistic Regression	0.730	0.787	0.790	0.791	0.794	0.821	0.855	0.883	0.896	0.927
SeenRank	0.626	0.673	0.728	0.751	0.746	0.779	0.806	0.839	0.869	0.892
Centroid	0.731	0.773	0.779	0.810	0.800	0.779	0.787	0.839	0.880	0.918
TextRank	0.373	0.398	0.485	0.540	0.624	0.664	0.714	0.728	0.764	0.783

FIGURE 5.6: Performance comparison of ranking techniques using NDCG scores.

Technique	@ 10	@ 20	@ 30	@ 40	@ 50	@ 60	@ 70	@ 80	@ 90	@ 100
TwitterEventInfoRank	100%	100%	100%	100%	100%	100%	100%	97.5%	96.6%	96.0%
LexRank	90.0%	80.0%	76.6%	65.0%	64.0%	63.3%	60.0%	62.5%	64.4%	64.0%
RTRank	80.0%	85.0%	86.6%	85.0%	86.0%	88.3%	90.0%	91.3%	92.2%	90.0%
Logistic Regression	60.0%	75.0%	76.6%	72.5%	74.0%	71.6%	68.5%	71.3%	71.1%	73.0%
SeenRank	80.0%	85.0%	80.0%	75.0%	72.0%	68.3%	70.0%	67.5%	65.5%	64.0%
Centroid	60.0%	60.0%	60.0%	62.5%	64.0%	66.6%	67.1%	67.5%	70.0%	68.0%
TextRank	0.00%	10.0%	13.3%	25.0%	28.0%	35.0%	42.8%	45.0%	47.8%	51.0%

FIGURE 5.7: Performance comparison of ranking techniques using precision scores.

Technique	@ 10	@ 20	@ 30	@ 40	@ 50	@ 60	@ 70	@ 80	@ 90	@ 100
TwitterEventInfoRank	100%	100%	100%	97.5%	98%	96.7%	95.7%	95.0%	95.5%	96.0%
LexRank	80.0%	85.0%	76.6%	72.5%	76.0%	78.3%	72.8%	73.7%	73.3%	74.0%
RTRank	60.0%	70.0%	76.6%	75.0%	70.0%	71.6%	71.4%	75.0%	73.3%	69.0%
Logistic Regression	100%	100%	100%	97.5%	96.0%	91.6%	92.8%	93.7%	93.3%	92.0%
SeenRank	70.0%	65.0%	70.0%	67.5%	62.0%	61.6%	57.1%	57.5%	55.5%	55.0%
Centroid	70.0%	75.0%	76.7%	82.5%	78.0%	71.6%	65.7%	66.3%	66.7%	66.0%
TextRank	10.0%	5.00%	13.3%	15.0%	22.0%	21.6%	24.3%	21.3%	22.2%	21.0%

FIGURE 5.8: Performance comparison of ranking techniques using precision scores.

selected. The choice of the time period was made on the basis of the interesecton of the time period of the tweets collected by us and that provided by Seen for the same event. There were 21641 tweets for Millions March NYC and 37429 tweets for Sydney Siege, respectively. We obtained the ranked tweets for all the seven approaches. For all the approaches except *SeenRank* the tweets were sorted in decreasing order on the basis of the ranking scores as the primary key and time of posting as the secondary key. This was done in order to get the most informative yet recent tweets at the top of the order. For *SeenRank* we sorted the tweets in terms of the scores assigned to them by Seen, as showing recent informative tweets for an event is one of the features of their platform.

We then followed a standard user evaluation approach to judge the event-specific informativeness of ranked tweets and also the hashtags, text units, URLs, and users. A team of three independent annotators comprising of graduate students, having taken the course of Information Retrieval, were assigned the task of annotation. Necessary background of the events were given to the annotators along with suitable resources for learning more about the events.

Tweet Annotations: The ranked tweets were annotated on an event-specific informativeness-scale of 1 to 3 by the three independent annotators. We provide sample tweets for each of them taking the Sydney Siege event as our example. The value of 1 was assigned to tweets that does not contain any event related information (for e.g. *SteveSmith*

becomes Australias 45th Test captain http://t.co/nYh9DqRXxh #sydneysiege #Martin-Place Lindt #MYEFO #siege Ray Hadley Muslims ISIS). Value of 2 was assigned to tweets that were related to the event yet they did not provide useful event-specific information (for e.g. *RT @TheDavidStevens: It wasn't just the policeman grabbing that girl in his arms, it was every Australian watching on too #sydneysiege*). A value of 3 was assigned to tweets that not only provided useful event-specific informative content but also led the user to more detailed information following the URLs mentioned in the tweet (for e.g. *RT @FoxNews: MORE: Police confirm 3 hostages escape Sydney cafe, unknown number remain inside http://t.co/pcAt91LIdS #Sydneysiege*). The annotators assigned scores to top 100 tweets ranked according to each of the seven strategies. Thereafter, we computed *Inter Indexer Consistency* (IIC) values [50] for the annotations of the two datasets. The average IIC scores obtained are shown in Table **??**. The IIC values for both the events fall in the acceptable range of accuracy of annotations. A tweet might be assigned three different scores by the annotators. In that scenario we find the average of the three scores and round it off to the smallest positive integer and assign a single score to each tweet. We also report the total average scores for top 100 tweets for both the events in Table **??**.

Hashtags, Text Units and URL Annotations: A similar annotation strategy was taken for annotating the top 50 hashtags, text units and URLs obtained using TwitterEventInfoRank. For hashtags and text units the annotators were asked to look at the tweets that consisted them. If the tweets primarily led to event-specific informative content then a score of 3 was assigned. If the tweets led to related but not so informative content about the event then they were assigned a score of 2. Hashtags and text units that were irrelevant and did not lead to any event related content, were assigned a score of 1. Similarly, the annotators visited the links for each URL, and based on the content they assigned them a score between 1-3. If the URLs were videos and images, then they further visited the tweet containing them in order to understand the context and scored them accordingly. Table **??** shows their average IIC scores and total average scores for top 50 ranks.

User Annotations: For annotating users we selected 5 random tweets for each of the top 50 users ranked according to *TwitterEventInfoRank*. An user was assigned a score of 3 if more than three of his tweets out of five got a score of 3 in the event-specific informativeness scale as already explained earlier. If three of his tweets get a score of 3 then the user gets a score of 2. Otherwise, a score of 1 is assigned to the user. Table **??** shows average IIC scores and total average scores for top 50 users.

NDCG@n and Precision@n: After being assured about consistency and accuracy of annotations, we moved to compute the *Normalized Discounted Cumulative Gain* (NDCG) [51] and Precision [52] values at each of the hundred recall levels. The NDCG values consider both the position and event-specific informativeness scores of the tweets. The NDCG value up-to position p in the ranking is given by equation 10, where DCG_p denotes the *discounted cumulative gain up-to position p* and is calculated using equation 9, and $IDCG_p$ denotes the *ideal discounted cumulative gain* value till position p in the ranking, or in other words the maximum possible DCG_p value till position p. rel_i denotes the graded relevance of the result at position i. In the context of our evaluation rel_i represents the average rounded score in the scale of (1-3) that has been assigned by the annotators to the tweet at position i in the ranked list of top 100 tweets.

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(i+1)} \tag{5.13}$$

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{5.14}$$

Precision@n is measured using equation 11. A tweet was considered to be relevant if it has a score of either 3 or 2 and was considered irrelevant if it has a score of 1.

$$\frac{No.\,of\,relevant\,tweets\,at\,position\,n}{n} \tag{5.15}$$

5.8.4.3 Event Analytics

Top Five Event-specific Informative Hashtags for Sydney Siege Event

1. #sydneysiege
2. #SydneySiege
3. #Sydneysiege
4. #MartinPlace
5. #9News

Top Five Event-specific Informative Text Units for Sydney Siege Event

1. police

2. sydney
3. reporter
4. lindt
5. isis

Top Five Event-specific Informative URLs for Sydney Siege Event

1. http://www.cnn.com/2014/12/15/world/asia/australia-sydney-hostage-situation/index.html
2. http://www.bbc.co.uk/news/world-australia-30474089
3. http://edition.cnn.com/2014/12/15/world/asia/australia-sydney-siege-scene/index.html
4. http://rt.com/news/214399-sydney-hostages-islamists-updates/
5. http://www.newsroompost.com/138766/sydney-cafe-siege-ends-gunman-among-two-killed

Top Five Event-specific Informative Tweet Excerpts for Sydney Siege Event

1. RT @faithcnn: Hostage taker in Sydney cafe has demanded 2 things: ISIS flag and; phone call with Australia PM Tony Abbott #SydneySiege http://t.co/a2vgrn30Xh
2. Aussie grand mufti and; Imam Council condemn #Sydneysiege hostage capture http://t.co/ED98YKMxqM - LIVE UPDATES http://t.c...
3. RT @PatDollard: #SydneySiege: Hostages Held By Jihadis In Australian Cafe - WATCH LIVE VIDEO COVERAGE http://t.co/uGxmd7zLpc #tcot #pjnet sydney-siege-scene/index.html
4. RT @FoxNews: MORE: Police confirm 3 hostages escape Sydney cafe, unknown number remain inside http://t.co/pcAt91LIdS #Sydneysiege
5. Watch #sydneysiege police conference live as hostages are still being held inside a central Sydney cafe http://t.co/OjulBqM7w2 #c4news

Three Randomly Selected Tweets for Top Three Event-specific Informative Users posting about Sydney Siege Event.

1. **User 1**
   (a) RT @cnni: Hostage taker in Sydney cafe demands ISIS flag and call with Australian PM, Sky News reports. http://t.co/a2vgrn30Xh #sydneysiege
   (b) RT @DR_SHAHID: Hostage taker demands delivery of an #ISIS flag and a conversation with Prime Minister Tony Abbott http://t.co/xTSDMKCPcD
   (c) RT @SkyNewsBreak: Update - New South Wales police commissioner confirms five hostages have escaped from the Lindt cafe in Sydney #sydneysiege
2. **User 2**
   (a) RT @smh: NSW Police Deputy Commissioner Catherine Burn will hold a press conference to update on the #SydneySiege at 6.30pm.
   (b) RT @Y7News: Helpful travel advice for commuters heading out of #Sydney's CBD this evening - http://t.co/aQx2lvSosm #sydneysiege
   (c) RT @hughwhitfeld: British PM David Cameron informed of #sydneysiege ..UK Foreign Office is in touch with Aus authorities
3. **User 3**
   (a) RT @RT_com: #SYDNEY: Gunman tall man in late 40s, dressed in black – eyewitness http://t.co/m51P8dUPhB #SydneySiege http://t.co/NvJzFsGrFN
   (b) RT @NewsAustralia: 2GB's Ray Hadley claims hostage takers in #SydneySiege "wants to speak to Prime Minister Abbott live on radio."
   (c) RT @BBCWorld: "Profoundly shocking" -Australia PM Tony Abbott delivers second #sydneysiege statement. MORE: http://t.co/VaKt3ZpRZR

Top Five Event-specific Informative Hashtags for Millions March NYC Event

1. #MillionsMarchNYC
2. #BlackLivesMatter
3. #ICantBreathe
4. #ShutItDown
5. #millionsmarchnyc

Top Five Event-specific Informative Text Units for Millions March NYC Event

1. police
2. nyc
3. eric
4. protesters
5. nypd

Top Five Event-specific Informative URLs for Millions March NYC Event

1. http://rt.com/usa/214203-protests-police-brutality-nationwide/index.html
2. http://mashable.com/2014/12/13/time-lapse-new-york-protest-march/?utm_cid=mash-com-Tw-main-link
3. http://www.cbsnews.com/news/eric-garner-ferguson-missouri-protesters-converge-on-washington/
4. http://www.huffingtonpost.com/2014/12/13/millions-march-nyc_n_6320348.html?ncid=tweetlnku
5. https://www.youtube.com/watch?v=Iz7hkfNmfTY&feature=youtu.be

Top Five Event-specific Informative Tweet Excerpts for Millions March NYC Event

1. RT @rightnowio_feed: Timelapse video reveals massive size of New York City prot... http://t.co/oHtIhEK969 #Soho #Millionsmarchnyc #NEWYorkC..
2. "@Breaking911: BREAKING NOW: #NYPD OFFICER INJURED ON THE BROOKLYN BRIDGE BY PROTESTERS THROWING ITEMS AT OFFICERS #MillionsMarchNYC" Great
3. RT @mohkeit: MT @WSJ: march to NYPD headquarters to protest police brutality #MillionsMarchNYC http://t.co/zhNSngjbkN http://t.co/YLMJ8uJnJ
4. RT @NaomiCampbell: Peaceful March Saturday Dec 13th Washington Square Park NYC 2:00pm march Tell everyone U know #MillionsMarchNYC
5. RT @anregarret: Incredible day! #MillionsMarchNYC On NYPD Headquarters To Protest Police Killings http://t.co/P2QHvxl9xb via @blackvoices

Three Randomly Selected Tweets for Top Three Event-specific Informative Users posting about Millions March NYC Event for a particular hour.

1. **User 1**
   (a) RT @mashable: Timelapse video reveals massive size of New York City protests http://t.co/zhqHpkDLk1 #MillionsMarchNYC http://t.co/WktxssAfDp
   (b) RT @DahmPublishing: RT@wendycarrillo: Real thugs wear flag pics and Eric Garner's eyes are haunting image #MillionsMarchNYC http://t.co/7wY...
   (c) RT @TheRoot: RT @mfmartinez: Protesters continue gathering in Washington Square Park #MillionsMarchNYC #TheRootMOW http://t.co/IwkQG1KjFg
2. **User 2**
   (a) RT @roqchams: Thousands march on NYPD headquarters to protest police terrorism http://t.co/yVyUVYkd9X http://t.co/X4QZrfOISh #MillionsMarchNYC
   (b) RT @NYjusticeleague: Hundreds killed. Ten Demands. One Continued Fight. Sign our petition at: http://t.co/KETNo6bS0V #MillionsMarchNYC htt...
   (c) RT @cobismith: Union Square now with NYPD in foreground, #MillionsMarchNYC protesters at right and; US national debt ticker on the left http:/...
3. **User 3**
   (a) RT @mashable: Timelapse video reveals massive size of New York City protests http://t.co/zhqHpkDLk1 #MillionsMarchNYC http://t.co/WktxssAfDp
   (b) RT @KeeganNYC: LOTS of NYPD waiting for protesters on the BK side of the Brooklyn Bridge #MillionsMarchNYC #ShutItDown #ICantBreathe http:/...
   (c) RT @Zegota42: . @KeeganNYC Protesters on Brooklyn Bridge leaving Manhattan Skyline behind. #MillionsMarchNYC #ICantBreathe http://t.co/UPvN...

NDCG@n and Precision@n values were calculated for all the seven approaches for each of the datasets. Figures 5.5 and 5.6 shows the NDCG curves for all the seven approaches on the Millions March NYC and the Sydney Siege events, respectively, for up-to 20 recall levels. Tables **??** and **??** presents the NDCG@n values and Precision@n values for different recall levels upto 100. It is quite evident from the figures and the tables that TwitterEventInfoRank approach outperforms all the baselines including the state-of-the-art approach of *SeenRank* in gaining event-specific information.

5.8.5 Discussion

On considering only the top 10 tweets we observed a substantial information gain of our algorithm over the state-of-the-art (*SeenRank*) and the baseline that performed second best for both the events. On comparing the values of NDCG@10 for the two events we found that our algorithm performs 13.96% (Millions March NYC) and 34.07% (Sydney Siege) better than the second best baseline technique, in identifying event-specific informative tweets. When compared with *SeenRank*, our algorithm was 64.53% (Millions March NYC) and 56.59% (Sydney Siege) better.

We also reasoned about the poor performance of *TextRank* in both the events. Since *TextRank* allowed random walks between homogeneous nodes, the strong association of non-informative nodes with the informative ones might have lowered the final scores of the informative nodes. The strong association of non-informative nodes with informative ones can be attributed to the spamming activity as already explained earlier in the paper. This also proves that our framework is robust against spams and is very effective in retrieving the most informative content related to events from the noisy stream of tweets in Twitter.

We also show the top 5 event-specific informative hashtags, text units, URLs and tweets for the Sydney Siege event in Table **??**, and three randomly selected tweets for top three event-specific informative users for the same event in Table **??**. Due to space constraints we do not present similar tables for Millions March NYC.

5.9 Conclusion and Future Work

In this paper we studied the characteristics of informative and non-informative content produced in 3.8 million (approx) tweets during three real-life events in Twitter. From our analysis we obtained cues for identifying informative content. We also observed that the supervised models used for assigning informativeness scores to tweets are generic and are not always well suited for identifying event-specific informative tweets. Moreover, they don't have the ability to simultaneously identify event-specific informative hashtags, text units, URLs and users. We were intrigued by the need of a model that identifies and ranks event-specific informative content from Twitter.

Using the cues from our analysis we found that hashtags used for annotating tweets, text units used for expressing the tweet content, URLs shared for providing additional information and the users posting them during an event are the main units of information

that could be leveraged for measuring event-specific informativeness. We identified mutually reinforcing relationships between the tweets, hashtags, text units, URLs and users posting them during an event, and represented their associations in a graph structure that forms the underlying framework for our ranking algorithm. We named the graph as *TwitterEventInfoGraph*. We also defined and quantified the semantics of the relationships between the vertices of the graph. Initial event-specific scores were assigned to the vertices. We proposed an algorithm *TwitterEventInfoRank* for ranking the vertices. The algorithm makes use of the mutually reinforcing chains formed between the vertices of *TwitterEventInfoGraph* for propagating the event-specific scores of a vertex to its neighboring vertices. The accumulated score of the vertices after the convergence of the algorithm is used for simultaneously rank streams of tweets, hashtags, text units, URLs and users producing them during two real-life events in terms of their event-specific informativeness. We obtained promising results using our proposed framework. The results were evaluated by comparing the performance of our approach with six other approaches including the state-of-the-art *SeenRank* algorithm used by Seen for ranking tweets displayed in their website. Our approach outperforms all the baselines by large margins for NDCG@n and Precision@n scores proving it to be the most effective and robust algorithm for identifying event-specific informative content from noisy stream of tweets in Twitter.

In this work we solved the problem of discovering event-specific informative content in tweets by proposing a robust and scalable framework. We pointed the ability of the framework to scale in a distributed processing environment. Our next step would be to extend the developed framework and implement it in a distributed computing environment, particularly integrating it with mapreduce. We also plan to use our framework for generating event summaries from Twitter, implementing event-centric recommendation in microblog environment and create event identities from the ranked information units for identifying event related content in other social media channels.

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' [53] to international sporting events like the 'Winter Olympics' [54] to socio-political [18] and socio-economical [55] events that shook the world such as presidential elections [56], 'Egyptian Revolution' [57], 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 TweetTracker[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 [58]. Social media is being used for tracking terrorism activities [59], collective actions [60], 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 #Heatvs-Cavs 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. #Heatvs-Cavs #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

Literature Review

7.1 Event Identification in News Text

7.2 Event Identification in Social Media

7.3 Information Quality in Social Media

7.4 Ranking and Summarization of Short Textual Social Media Posts

7.5 Reference Tracking and Entity Resolution

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] S.D. Reese et al. Mapping the blogosphere professional and citizen-based media in the global news arena. *Journalism*, 8(3):235–261, 2007.

[8] N. Hamdy et al. Framing the egyptian uprising in arabic language newspapers and social media. *Journal of Communication*, 2012.

[9] T.J. Johnson et al. Wag the blog: How reliance on traditional media and the internet influence credibility perceptions of weblogs among blog users. *Journalism & Mass Communication Quarterly*, 81(3):622–642, 2004.

[10] Nitin Agarwal, Merlyna Lim, and Rolf T Wigand. Finding her master's voice: the power of collective action among female muslim bloggers. In *ECIS*, 2011.

[11] Nitin Agarwal, Merlyna Lim, and Rolf T Wigand. Online collective action and the role of social media in mobilizing opinions: A case study on women's right-to-drive campaigns in saudi arabia. In *Web 2.0 Technologies and Democratic Governance*, pages 99–123. Springer, 2012.

[12] Nitin Agarwal, Merlyna Lim, and Rolf T Wigand. Raising and rising voices in social media. *Business & Information Systems Engineering*, 4(3):113–126, 2012.

[13] L.A. Adamic et al. Power-law distribution of the world wide web. *Science*, 287 (5461):2115–2115, 2000.

[14] LOmariba. Is new media posing a serious challenge to traditional media?". Technical report, University of Westminster, 2009.

[15] Z. Harb. Arab revolutions and the social media effect. *M/C Journal*, 14(2), 2011.

[16] C. Anderson. *Long Tail, The, Revised and Updated Edition: Why the Future of Business is Selling Less of More*. Hyperion, 2008.

[17] A. Younus et al. What do the average twitterers say: A twitter model for public opinion analysis in the face of major political events. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 618–623. IEEE, 2011.

[18] V.K. Singh et al. 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.

[19] S. Vieweg et al. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1079–1088. ACM, 2010.

[20] F. Cheong et al. Social media data mining: A social network analysis of tweets during the 2010–2011 australian floods. *PACIS 2011 Proceedings*, 2011.

[21] A. Marcus et al. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 227–236. ACM, 2011.

[22] A.M. Popescu et al. Extracting events and event descriptions from twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 105–106. ACM, 2011.

[23] Raphaël Troncy et al. Linking events with media. In *Proceedings of the 6th International Conference on Semantic Systems*, page 42. ACM, 2010.

[24] T. Rattenbury et al. Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110. ACM, 2007.

[25] Hila Becker, Mor Naaman, and Luis Gravano. Selecting quality twitter content for events. *ICWSM*, 11, 2011.

[26] Manish Gupta, Peixiang Zhao, and Jiawei Han. Evaluating event credibility on twitter. In *SDM*, pages 153–164. SIAM, 2012.

[27] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2451–2460. ACM, 2012.

[28] S. Brin et al. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[29] T.H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4): 784–796, 2003.

[30] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[31] J. Ramos. Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, 2003.

[32] G.H. Golub et al. *Matrix computations*, volume 3. Johns Hopkins University Press, 1996.

[33] A.N. Langville et al. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380, 2004.

[34] B. Ekdale et al. From expression to influence: Understanding the change in blogger motivations over the blogspan. *AEJMC, Washington, DC*, 2007.

[35] 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.

[36] 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.

[37] 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.

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

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

[40] 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.

[41] Kalina Bontcheva, Genevieve Gorrell, and Bridgette Wessels. Social media and information overload: Survey results. *arXiv preprint arXiv:1306.0813*, 2013.

[42] Manuel Gomez Rodriguez, Krishna Gummadi, and Bernhard Schoelkopf. Quantifying information overload in social media and its impact on social contagions. *arXiv preprint arXiv:1403.6838*, 2014.

[43] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics, 2009.

[44] 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.

[45] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM" 14)*, number EPFL-CONF-203561, 2014.

[46] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd international conference on World Wide Web companion*,

pages 1021–1024. International World Wide Web Conferences Steering Committee, 2013.

[47] Sarita Yardi, Daniel Romero, Grant Schoenebeck, et al. Detecting spam in a twitter network. *First Monday*, 15(1), 2009.

[48] Hila Becker, Mor Naaman, and Luis Gravano. Beyond trending topics: Real-world event identification on twitter. *ICWSM*, 11:438–441, 2011.

[49] Jimmy Lin and Michael Schatz. Design patterns for efficient graph algorithms in mapreduce. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, pages 78–85. ACM, 2010.

[50] L Rolling. Indexing consistency, quality and efficiency. *Information Processing & Management*, 17(2):69–76, 1981.

[51] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[52] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

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

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

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

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

[57] 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.

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

[59] 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.

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