University of Arkansas at Little Rock

Doctoral Thesis

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

Author:
Debanjan Mahata

Supervisor:
Dr. John R. Talburt

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

in

Integrated Computing
Information Quality Track
Department of Information Science

April 2015

Declaration of Authorship

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

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

Signed: ________________________________

Date: ________________________________

"Torture the data, and it will confess to anything."

Ronald Coase, Economics, Nobel Prize Laureate

Abstract

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

Acknowledgements

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

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

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

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

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

Contents

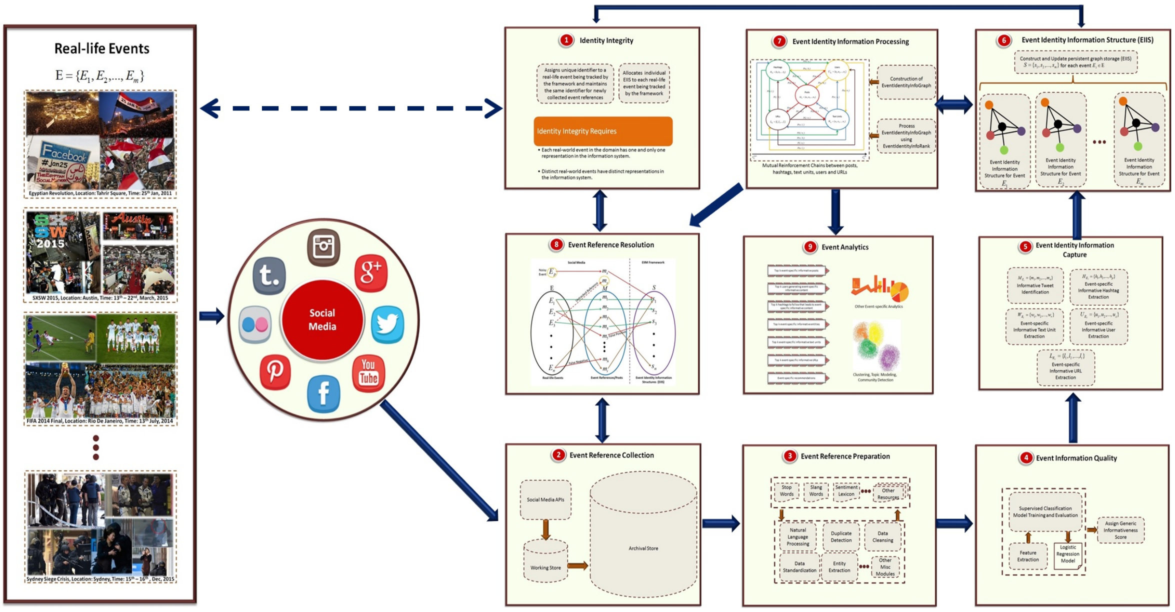
List of Figures

List of Tables

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

Dissertation Overview

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


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

Related Filed Patent

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

Related Publications

- **Debanjan Mahata**, John R. Talburt and Vivek Kumar Singh; *Identifying and Ranking of Event-specific Entity-centric Informative Content from Twitter.* 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 Background : Entity Identity Information Management in Master Data Management

1.3 Problem Definition and Research Questions

1.4 General Challenges in Mining Social Media Text

1.4.1 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] http://www.statisticbrain.com/twitter-statistics/

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

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

1.4.2 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 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.

1.4.3 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.

[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

1.4.4 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 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.

1.4.5 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.

1.4.6 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

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

error or perception in annotating the data. This might lead to wrong assumptions and false results.

1.5 Research Methodology

1.6 Research Contributions

1.7 Structure of the Thesis

Chapter 2

Literature Review

2.1 Event Identification in News Text

The event detection task [7] in the TDT program (Topic Detection and Tracking), led to significant advancements in the field of event-based organization of broadcast news. Some of the efforts in the TDT program focused on online event detection from continuous and real-time streams of textual news documents in newswires [8, 9]. While others explored the detection of past events from archived news documents [10].

The textual content in news documents are different from the short informal text common in the realm of social media. Most of these documents contain formal text with well-formed grammatical structures, enabling the researchers to rely on the state-of-the-art natural language processing techniques. Named entity extraction and Parts-of-Speech (POS) tagging are among the widely used techniques. Zhang et al. [11] extracted named entities and POS tags from textual news documents, and used them to reweigh tf-idf representations of these documents for the new event detection task. Filatova and Hatzivassiloglou [12] identified named entities corresponding to participants, locations, and times in text documents, and then used the relationships between certain types of entity pairs to detect event content. Hatzivassiloglou et al. [13] used linguistic features (e.g., noun phrase heads, proper names) and learned a logistic regression model for combining these features into a single similarity value. Makkonen et al. [14] extracted meaningful semantic features such as names, time references, and locations, and learned a similarity function that combines these metrics into a single clustering solution. They concluded that augmenting documents with semantic terms did not improve performance, and reasoned that inadequate similarity functions were partially to blame.

Extracting events from text has been the focus of numerous studies as part of the NIST initiative for Automatic Content Extraction (ACE) [15, 16]. The ACE program defines event extraction as a supervised task, given a small set of predefined event categories and entities, with the goal of extracting a unified representation of the event from text via attributes (e.g., type, subtype, modality, polarity) and event roles (e.g., person, place, buyer, seller). Ahn [15] divided the event extraction task into different subtasks, including identification of event keyword triggers (see Chapter 2), and determination of event coreference, and then used machine learning methods to optimize and evaluate the results of each subtask. Ji and Grishman [16] proposed techniques for extracting event content from multiple topically similar documents, instead of the traditional approach of extracting events from individual documents in isolation. In contrast with the predefined templates outlined by ACE, Filatova et al. [17] presented techniques to automatically create templates for event types, referred to as domains, given a set of domain instances (i.e., documents containing information related to events that belong to the domain).

As already discussed, social media documents are extremely concise, noisy and lacks well-established grammatical structures. Therefore, the techniques used in these works are not suitable for identification of events from social media. 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]. Therefore, new approaches had to be taken, leading to new techniques for detecting events in social media, which we discuss next.

2.2 Event Identification in Social Media

While event detection in textual news documents has been studied in depth, the identification of events in social media sites is still in its infancy. Several related papers explored the unknown event identification scenario in social media. Weng and Lee [18] proposed wavelet-based signal detection techniques for identifying real-life events on Twitter. These techniques can detect significant bursts or trends in a Twitter data stream but, unlike our work in Chapter 4, they do not filter the vast amount of non-event content that exists on Twitter. This, unfortunately, results in poor performance, with very low precision scores compared with the precision achieved by our methods. Related to our work in Chapters 4 and 5, Sankaranarayanan et al. [19] identified late breaking news events on Twitter using clustering, along with a text-based classifier and a set of news seeders, which are handpicked users known for publishing news (e.g., news agency feeds). As we discussed, such text-based and seeder-driven filtering of non-event data can be used to generate the event document stream we use in Chapter 5. Petrovic et al. [20] used locality-sensitive hashing to detect the first tweet associated with an event in a stream of Twitter messages. We use the general text-based classifier suggested in [19] and a method for identifying top events suggested by Petrovic et al. [20] as baseline approaches in our evaluation of the unknown identification methods of Chapter 4. While our work in the unknown event identification scenario focuses on timely, online, analysis, several efforts tried to address this task using retrospective analysis. Rattenbury et al. [21] analyzed the temporal usage distribution of tags to identify tags that correspond to events. Chen and Roy [22] used the time and location associated with Flickr image tags to discover event-related tags with significant distribution patterns (e.g.bursts) in both of these dimensions.

Recent efforts proposed techniques for known identification of events in social media. Many of these techniques rely on a set of manually selected terms to retrieve event-related documents from a single social media site [23, 24]. Sakaki et al. [23] developed techniques for identifying earthquake events on Twitter by monitoring keyword triggers (e.g., earthquake or shaking). In their setting, the type of event must be known a priori, and should be easily represented using simple keyword queries. Most related to our work in Chapter 6, Benson et al. [25] identified Twitter messages for concert events using statistical models to automatically tag artist and venue terms in Twitter messages. Their approach is novel and fully automatic, but it limits the set of identified messages for concert events to those with explicit artist and venue mentions. Importantly, both of these approaches are tailored to one specific social media site. In contrast, we propose methods for identifying social media documents across multiple sites with varying types of documents (e.g., photos, videos, textual messages). Our goal is to automatically retrieve social media documents for any planned event, without any assumption about the textual content of the event or its associated documents. While not exclusively in the social media domain, Tsagkias et al. [26] extracted named entities and quotations from news articles, as well as explicit links between news and social media documents, to identify social media utterances related to individual news stories. In contrast with their well formed, lengthy textual documents and explicitly linked content, content in our known event identification setting (Chapter 6) is brief and often noisy, and generally does not contain explicit links to social media documents.

2.3 Information Quality in Social Media

2.4 Ranking and Summarization of Short Textual Social Media Posts

There are many web hosted applications that supplements the default search provided by Twitter in order to effectively retrieve relevant and high quality tweets from different perspectives[1]. On going through these services we found that the most commonly used criteria for ranking tweets are recency, popularity based on retweets and favorite counts, authority of the users posting the tweets and content relevance. Twitter itself uses the popularity of the tweets and features mined from the profile of the users in order to provide personalized search results ordered by recency[2]. A study of different state-of-the-art features and approaches commonly used for ranking tweets has been documented by [27**?**]. Seen[3] is a new state-of-the-art platform that uses a proprietary algorithm named *SeenRank* for ranking event related tweet content for presenting event highlights and summaries. In this work, we consider *SeenRank* as one of our baselines. As the number of retweets of a tweet is widely used for ranking, we also use it as one of our baselines. In the context of our work we name the ranking scheme as *RTRank*

Apart from the existing real-world search applications, several adaptations of *PageRank* [28] has been proposed by the scientific community for ranking tweets and users in Twitter [29–31]. Various learning to rank approaches have been used for ordering tweets retrieved for a given query in terms of their relevance and quality [32**?** , 33]. None of these ranking techniques have been devised for event-specific content. An attempt to solve a similar problem presented in this paper was made by [34]. They represented tweets of an event in a cluster and calculated the similarity of individual tweets with the centroid of the cluster. Then they ranked the tweets based on the decreasing value of their similarity. We use this approach as one of our baselines.

2.5 Reference Tracking and Entity Resolution

[1] http://mashable.com/2009/04/22/twitter-search-services

[2] https://blog.twitter.com/2011/engineering-behind-twitter%E2%80%99s-new-search-experience

[3] http://seen.co

Chapter 3

Defining Events in Social Media

3.1 Events from Different Perspectives

3.1.1 Topic Detection and Tracking

3.1.2 Automatic Content Extraction

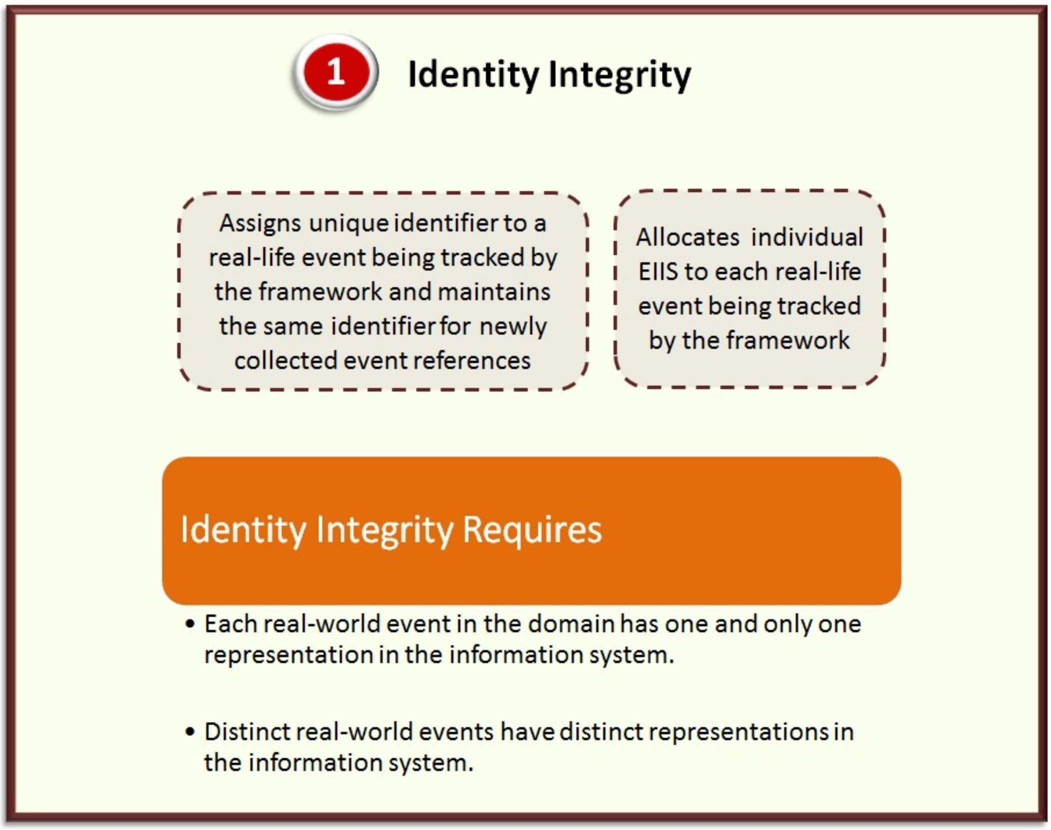
3.1.3 Multimedia Event Detection

3.2 Events in Social Media

Chapter 4

Event Identity Information Management (EIIM) Life Cycle for Social Media

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



Identity Integrity

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

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

Identity Integrity Requires

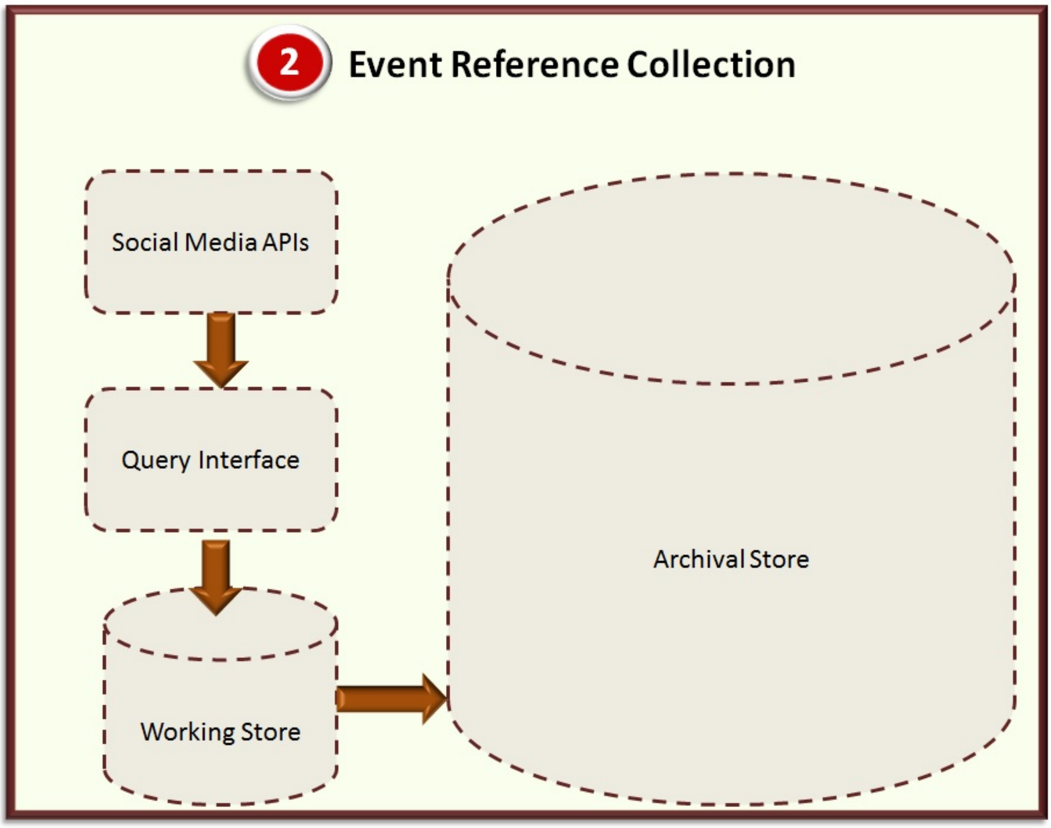
- Each real-world event in the domain has one and only one representation in the information system.
- Distinct real-world events have distinct representations in the information system.

4.1 Identity Integrity

One of the fundamental goals of the proposed framework is to maintain a one-to-one correspondence between real-world events being monitored and the Event Identity Information Structure (EIIS) of the corresponding events for ensuring identity integrity. Therefore, a separate EIIS is maintained corresponding to each event. As new events are introduced to the framework, a unique identifier is assigned to them along with the allocation of individual EIIS structures. The framework is expected to maintain the integrity throughout the EIIM life cycle, by consistently assigning the same identifier to the references of a tracked event. Modules of this component assigns 12 byte unique integers known as ObjectId to each event, and is also responsible for maintaining the same ObjectId for event ids of collected references and related EIIS. It is also the functionality of this component to assign the right identifier to the references resolved for an event by the Event Reference Resolution component.

4.2 Event Reference Collection

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



This component allows the framework to collect event references from different social media websites using its publicly available APIs (Application Programming Interface), and store them in the database after processing them using the next two components of the EIIM life cycle. Due to the semi-structured nature of the collected data, a NOSQL document oriented database management system (MongoDb) is used for storage. The choice of MongoDb was also driven by its ability to scale horizontally and perform operations on large volumes of data. For the experiments and analysis 4 million tweets (approx) related to five different events were collected using this component. Details of the collected event references are provided in Table 2. The tweets were collected over the given period of time, by providing a popular hashtag to the Twitter streaming API (for details about Twitter Data Collection please refer Appendix A.

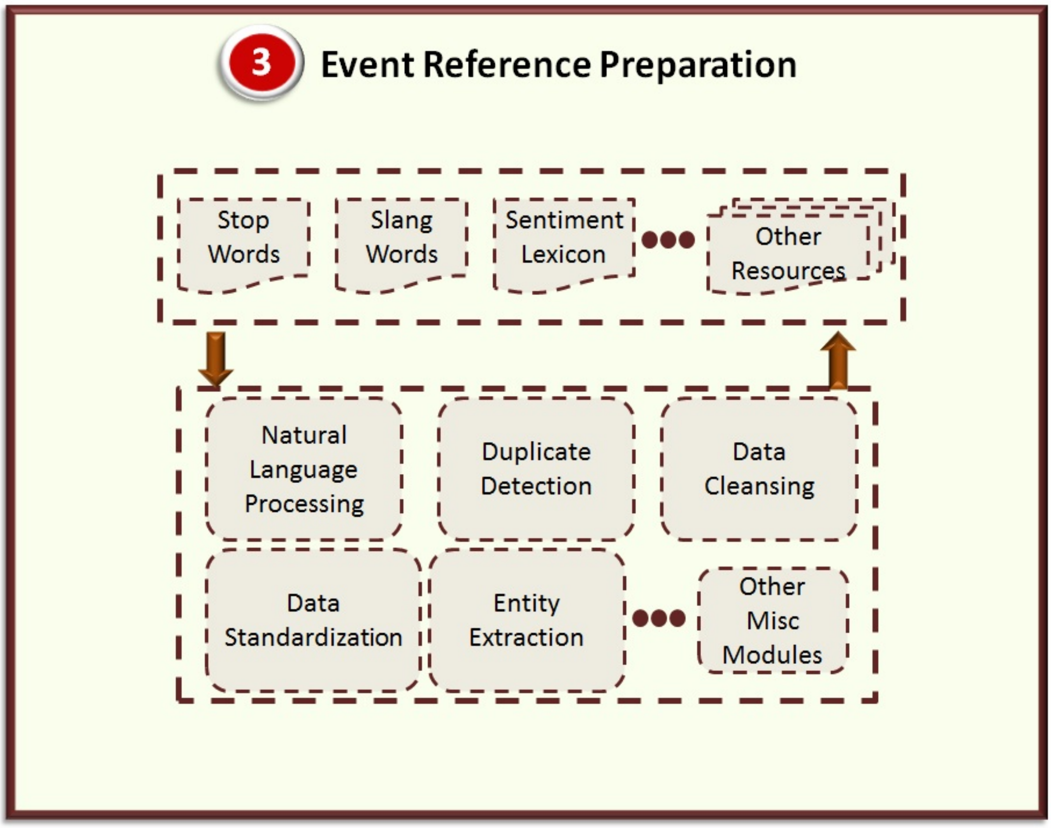
Table 4.1: Details of data collected for analyzing event related tweet content.

Event	**Query Hashtag**	**No. of Tweets**	**Time Period**
Sochi Winter Games 2014 ($http://goo.gl/sG4Rqd$)	#sochi2014	1958220	11th Feb,2014 to 3rd March, 2014
SXSW 2014 ($http://goo.gl/b6Nd6X$)	sxsw2014	1880557	8th March, 2014 to 16th March, 2014
CPAC 2014 ($http://goo.gl/9o1KUx$)	#cpac2014	18104	7th March, 2014 to 16th March, 2014
Millions March NYC ($http://goo.gl/I8WR4B$)	#millionsmarchnyc	56927	13th Dec, 2014 20:25:43 to 14th Dec, 2014 03:30:41
Sydney Siege ($http://goo.gl/qLguvG$)	#sydneysiege	398204	15th Dec, 2014 07:21:16 to 15th Dec, 2014 22:46:45

4.3 Event Reference Preparation

Preprocessing the raw references is an important stage of any data intensive application. This component performs a series of data preparation steps on the collected event tweets in order to make them suitable for further processing by the other components of the EIIM life cycle. It performs deduplication of tweets using md5 hashing scheme. Redundant copies of a tweet are filtered out keeping a single copy in the database. Parts-of-speech tagging is done using the default POS tagger available in the NLTK module. A standard list of English stop words is used for eliminating the stop words from the tweet

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



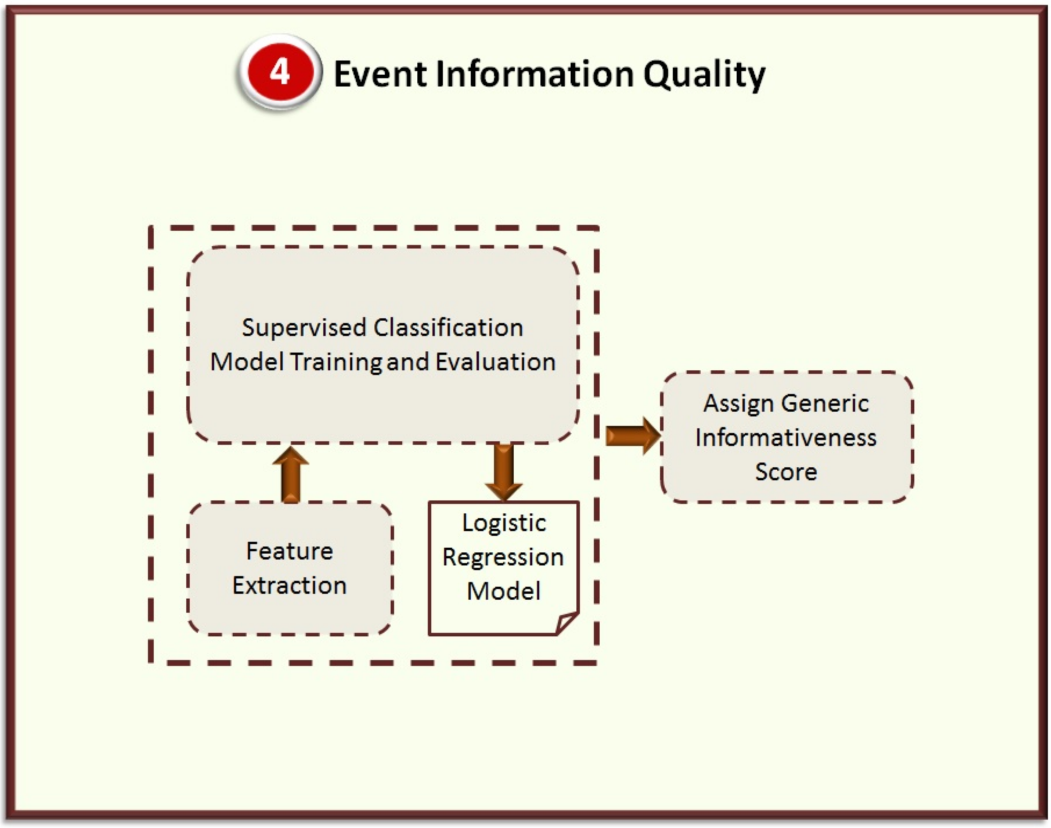
text. All the characters of a tweet are converted into lower case and special characters are removed. The tweets are tokenized into unigram tokens. User mentions, retweet symbol and URLs are removed during tokenization and are not considered as tokens.

A list of words expressing feelings in the internet, obtained from wefeelfine.org is used for detecting and extracting the feeling words from a tweet. Slang words commonly used in the internet and twitter specific slang publicly shared by FBI is combined together for compiling a list of English slang words. The modules use this list for detecting and extracting the slang words from the tweets, hashtags and text units. Retweet counts, favorite counts, verification information, user follower count, time information and expanded form of the URLs shared in the tweets are extracted from the metadata associated with each tweet, as retrieved using the Twitter API.

4.4 Event Information Quality

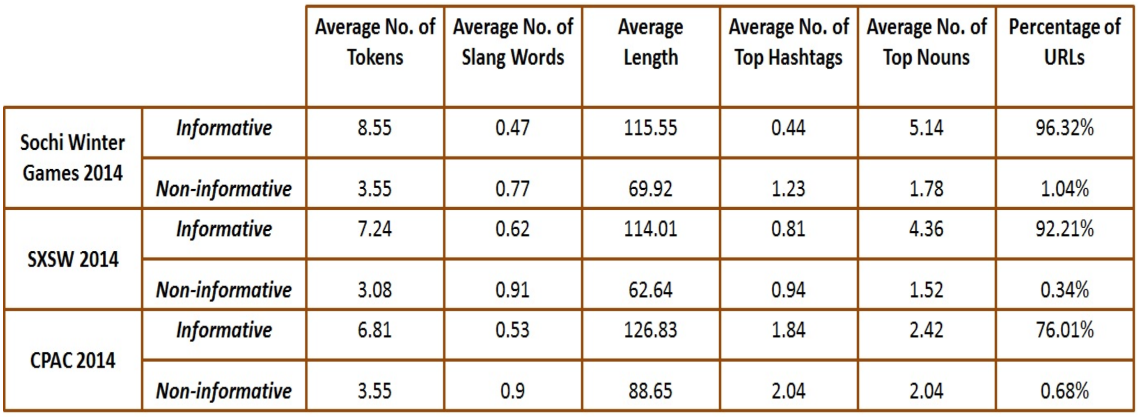
This component examines the quality of information present in the tweets collected for the events. It segregates the references having high likelihood of containing good quality event related information from the ones that are less likely to contain or point to good

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


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

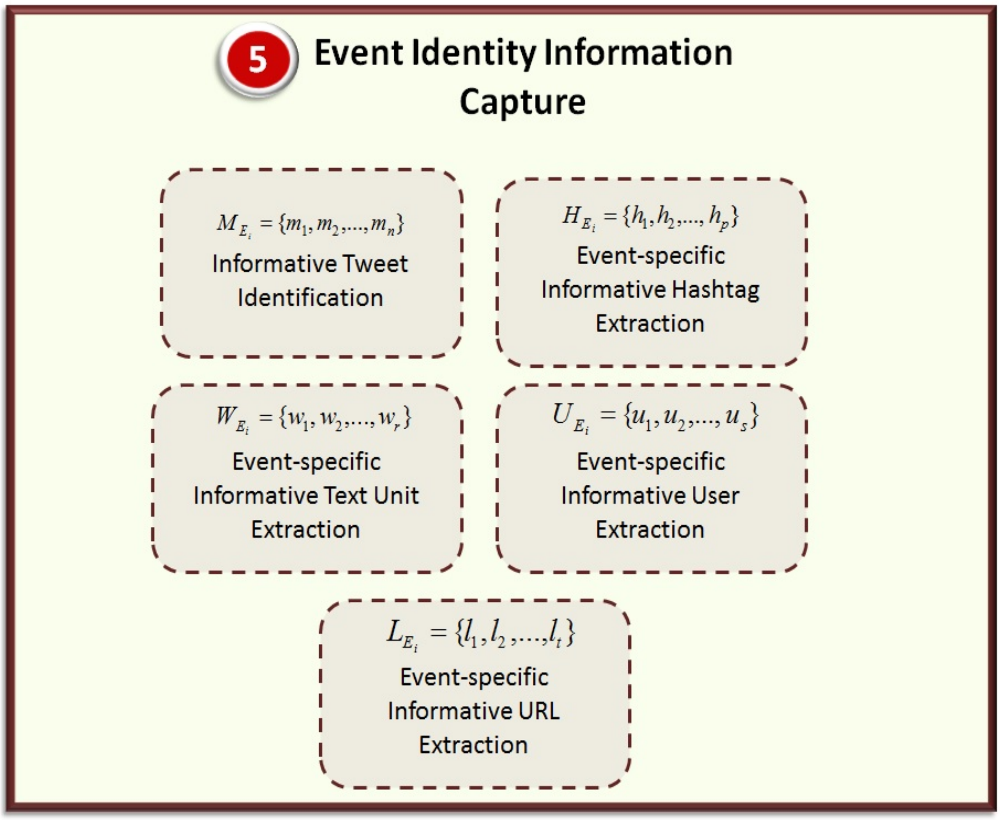
quality information. In order to make a generic module for identifying high quality event related informative references we implemented a logistic regression classifier trained on a publicly available annotated dataset provided by [28]. The tweets labeled as 'related and informative' were assigned a score of 1 and all the other tweets labeled as 'related-but not informative', and 'not related' were assigned a score of 0. Table 3 lists the features extracted from each tweet. The choice of features was governed by previous works related to identifying high quality information from Twitter as already pointed in the Related Work section. 10-fold cross validation was performed resulting in a model with an accuracy of 76.64The trained model is used for assigning a score between 0 (least informative) and 1 (most informative) to the tweets in real-time. Both the 'Event Reference Preparation' and the 'Event Information Quality' components work in collaboration with the 'Event Reference Collection' component in order to collect, prepare, assign quality score and store the tweets related to an event, obtained from Twitter streaming API, in real-time.

Figure 4.5: 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%

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



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

Table 4.2: 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 4.3: 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%		

4.5 Event Identity Information Capture

It is the component that aids in extracting event identity information units (explained later) from the already processed tweets and build the Event Identity Information Structure (EIIS) for an event. It also enables the framework to set a threshold between 0.0-1.0 for differentiating between high quality informative tweets from low quality non-informative ones related to an event. The event identity information units are then extracted from the high quality informative tweets. In order to understand what might consist of the event identity information units that would represent the EIIS, we conducted a detailed analysis of 3.8 million tweets collected for three events. Details of the data collected are provided in Table 6. The data collection task was accomplished by Event Reference Collection component and was then preprocessed by the Event Reference Preparation component.

The logistic regression model developed for the Event Information Quality component was used for assigning scores to all the 3.8 million tweets in the dataset. The tweets getting a score greater than 0.7 were considered as instances of high quality informative tweets. Those getting a score lesser than 0.3 were considered as instances of low quality non-informative tweets. Average values of different content characteristics of the tweets were calculated. Top ten percent of the frequently occurring hashtags and nouns were considered as top hashtags and top nouns respectively, for the analysis. Some of the characteristics that were prominently different for informative and non-informative tweets are listed in Table 5. As presented in the table, for all the three events, on an average the informative tweets are marked by a higher number of tokens per tweet and greater occurrence of top nouns. The average length of informative tweets is also more than the non-informative ones. The percentage of informative tweets having URLs is strikingly high. A greater use of slang words is observed in non-informative tweets. However, greater occurrence of top hashtags in non-informative tweets intrigued us to look into the content and obtain a detailed view of it. We observed that a lot of non-informative tweets have used popular hashtags with unrelated content and URLs directing to irrelevant information. This is typical of spam tweets as already reported by [30]. Although not shown due to space constraints, 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 average number of feeling words used in informative tweets were also relatively higher than the feeling words used in the non-informative tweets.

The above observations gave us an idea of how high quality informative content related to events is produced in Twitter and the characteristics that differentiate them from low quality non- informative content. It is now intuitive that the informative tweets are more expressive, formal and lengthier, marked by higher presence of nouns. The high presence of nouns indicates that these tweets also contain information about people, places, organizations, etc, associated with the events, which is vital information about any event and is ideal for representing its identity. Due to the limitations 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 can also be concluded by the fact that as they have more followers they are encouraged to share informative content. Conversely, since they share informative content they are followed by a large number of other users interested in the content shared by them. Based on the above analysis we decided to build the EIIS for an event composed of the following event identity information units:

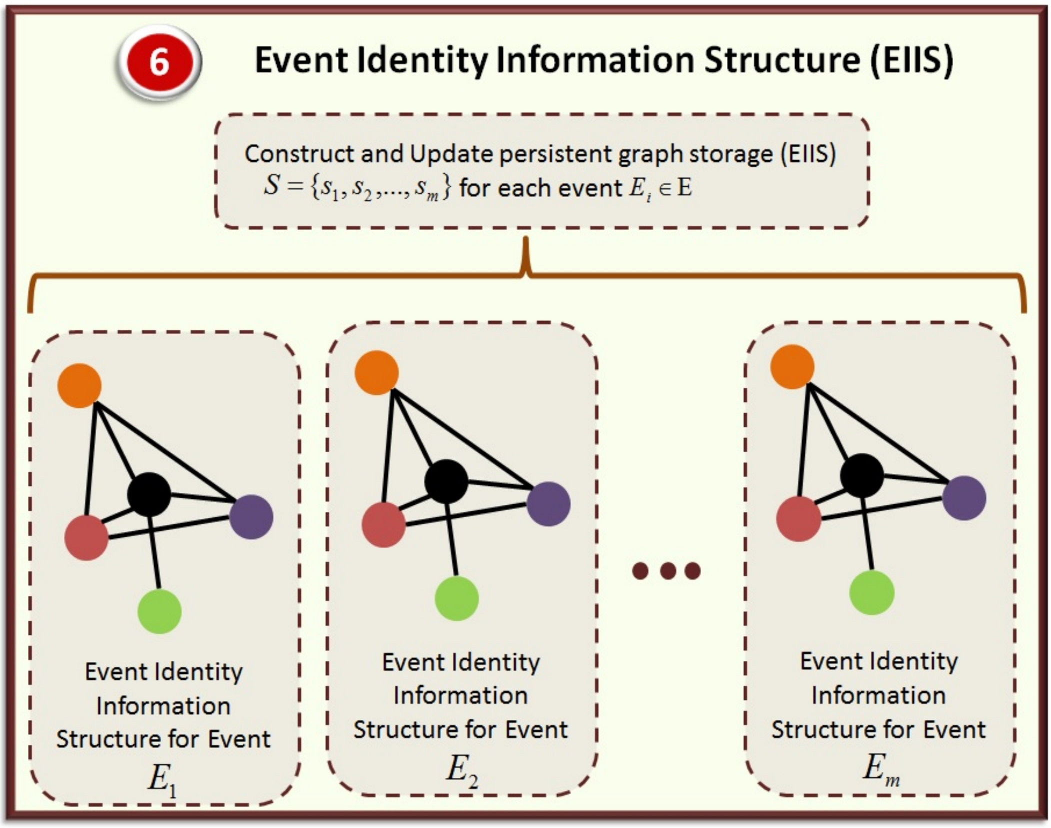
4.6 Event Identity Information Structure

4.7 Event Identity Information Processing

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.*

FIGURE 4.7: Event Identity Information Structure component of the EIIM life cycle.



6 Event Identity Information Structure (EIIS)

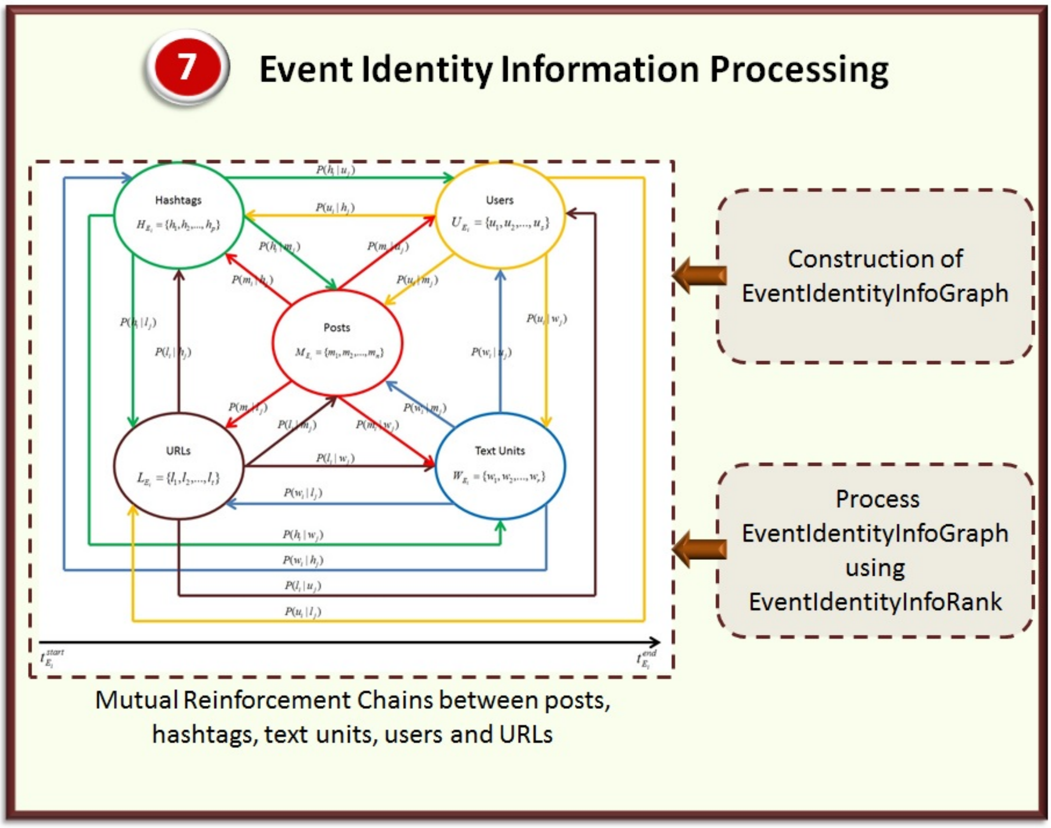
Construct and Update persistent graph storage (EIIS) $S = \{s_1, s_2, ..., s_m\}$ for each event $E_i \in \mathrm{E}$

Event Identity Information Structure for Event E_1

Event Identity Information Structure for Event E_2

Event Identity Information Structure for Event E_m

FIGURE 4.8: Event Identity Information Processing component of the EIIM life cycle.



7 Event Identity Information Processing

Mutual Reinforcement Chains between posts, hashtags, text units, users and URLs

Construction of EventIdentityInfoGraph

Process EventIdentityInfoGraph using EventIdentityInfoRank

Table 4.4: 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.

- 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,*

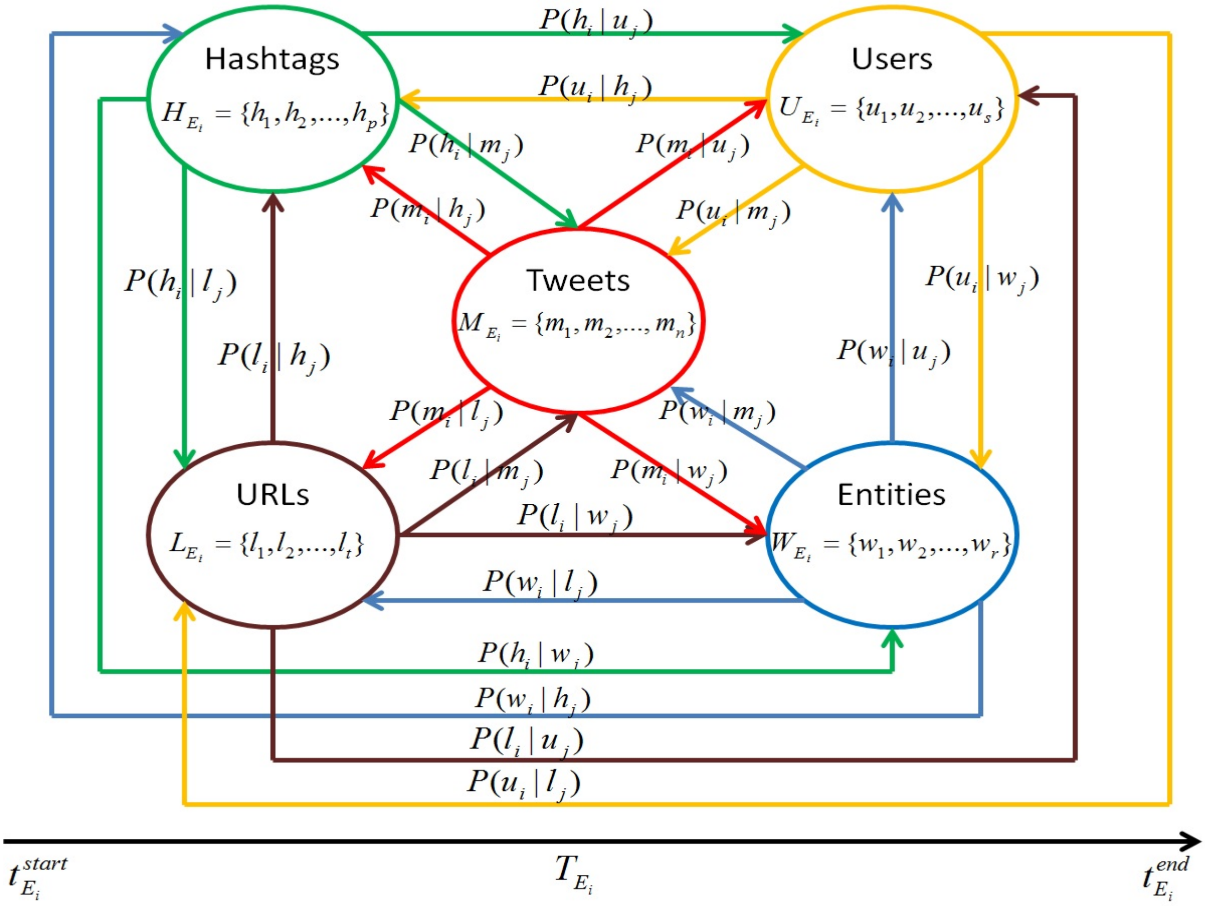
  (b) *event-specific informative hashtags,*

  (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:

FIGURE 4.9: Mutual Reinforcement Chains in Twitter for an event.



(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* [35] for the event E_i as shown in Figure 4.9. We represent this relationship in a graph $\mathbf{G} = (\mathbf{V},\mathbf{D})$, 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 4.4.

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*.

4.7.1 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 4.4. 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.

$$Score(h_i) = \frac{freq(h_i)}{max\{freq(h_1), freq(h_2), ..., freq(h_p)\}} \tag{4.1}$$

$$Score(w_i) = \frac{freq(w_i)}{max\{freq(w_1), freq(w_2), ..., freq(w_r)\}} \tag{4.2}$$

$$Score(u_i) = \frac{followers(u_i)}{max\{followers(u_1), ..., followers(u_r)\}} \tag{4.3}$$

$$Score(l_i) = \frac{freq(l_i)}{max\{freq(l_1), freq(l_2), ..., freq(l_r)\}} \tag{4.4}$$

The relationships between two different subsets of vertices in graph $\mathbf{G}$ is denoted by an affinity matrix. For e.g., $\mathbf{A_{E_i}^{MH}}$ denotes the $\mathbf{M_{E_i}} - \mathbf{H_{E_i}}$ affinity matrix for event E_i, where $\mathbf{(i, j)^{th}}$ 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 4.4. Similarly, $\mathbf{A_{E_i}^{WH}}$ denotes the $\mathbf{W_{E_i}} - \mathbf{H_{E_i}}$ 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:

$$R^{M(k+1)}_{E_i} = A^{MM(k)}_{E_i} R^{M(k)}_{E_i} + A^{MH(k)}_{E_i} R^{H(k)}_{E_i} + A^{MW(k)}_{E_i} R^{W(k)}_{E_i} + A^{MU(k)}_{E_i} R^{U(k)}_{E_i} + A^{ML(k)}_{E_i} R^{L(k)}_{E_i} \tag{4.5}$$

$$R^{H(k+1)}_{E_i} = A^{HM(k)}_{E_i} R^{M(k)}_{E_i} + A^{HH(k)}_{E_i} R^{H(k)}_{E_i} + A^{HW(k)}_{E_i} R^{W(k)}_{E_i} + A^{HU(k)}_{E_i} R^{U(k)}_{E_i} + A^{HL(k)}_{E_i} R^{L(k)}_{E_i} \tag{4.6}$$

$$R^{W(k+1)}_{E_i} = A^{WM(k)}_{E_i} R^{M(k)}_{E_i} + A^{WH(k)}_{E_i} R^{H(k)}_{E_i} + A^{WW(k)}_{E_i} R^{W(k)}_{E_i} + A^{WU(k)}_{E_i} R^{U(k)}_{E_i} + A^{WL(k)}_{E_i} R^{L(k)}_{E_i} \tag{4.7}$$

$$R^{U(k+1)}_{E_i} = A^{UM(k)}_{E_i} R^{M(k)}_{E_i} + A^{UH(k)}_{E_i} R^{H(k)}_{E_i} + A^{UW(k)}_{E_i} R^{W(k)}_{E_i} + A^{UU(k)}_{E_i} R^{U(k)}_{E_i} + A^{UL(k)}_{E_i} R^{L(k)}_{E_i} \tag{4.8}$$

$$R^{L(k+1)}_{E_i} = A^{LM(k)}_{E_i} R^{M(k)}_{E_i} + A^{LH(k)}_{E_i} R^{H(k)}_{E_i} + A^{LW(k)}_{E_i} R^{W(k)}_{E_i} + A^{LU(k)}_{E_i} R^{U(k)}_{E_i} + A^{LL(k)}_{E_i} R^{L(k)}_{E_i} \tag{4.9}$$

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

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

Let

$$R_{E_i} = \begin{pmatrix} R^M_{E_i} \\ R^H_{E_i} \\ R^W_{E_i} \\ R^U_{E_i} \\ R^L_{E_i} \end{pmatrix}$$

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

$$\Delta_{E_i}.R_{E_i} = \lambda.R_{E_i} \tag{4.10}$$

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

$$\overline{\Delta}_{E_i} = \alpha\hat{\Delta}_{E_i} + (1-\alpha)E \tag{4.11}$$

$$E = p \times [1]_{1\times k} \tag{4.12}$$

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 $\overline{\Delta}^2_{E_i}$ is greater than 0.

Following steps are taken next,

1. We initialize the rank vectors $(R^{M(0)}_{E_i}, R^{H(0)}_{E_i}, R^{W(0)}_{E_i}, R^{U(0)}_{E_i}, R^{L(0)}_{E_i})$ 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

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

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

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

```
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 $\Delta_{E_i}$;
Make matrix $\Delta_{E_i}$ 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 $\hat{M}_{E_i}$,$\hat{H}_{E_i}$,$\hat{W}_{E_i}$,$\hat{U}_{E_i}$,$\hat{L}_{E_i}$;
```

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 [36]. We plan to work on it in the future and implement our framework using hadoop and mapreduce environment.

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.

Table 4.5: 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

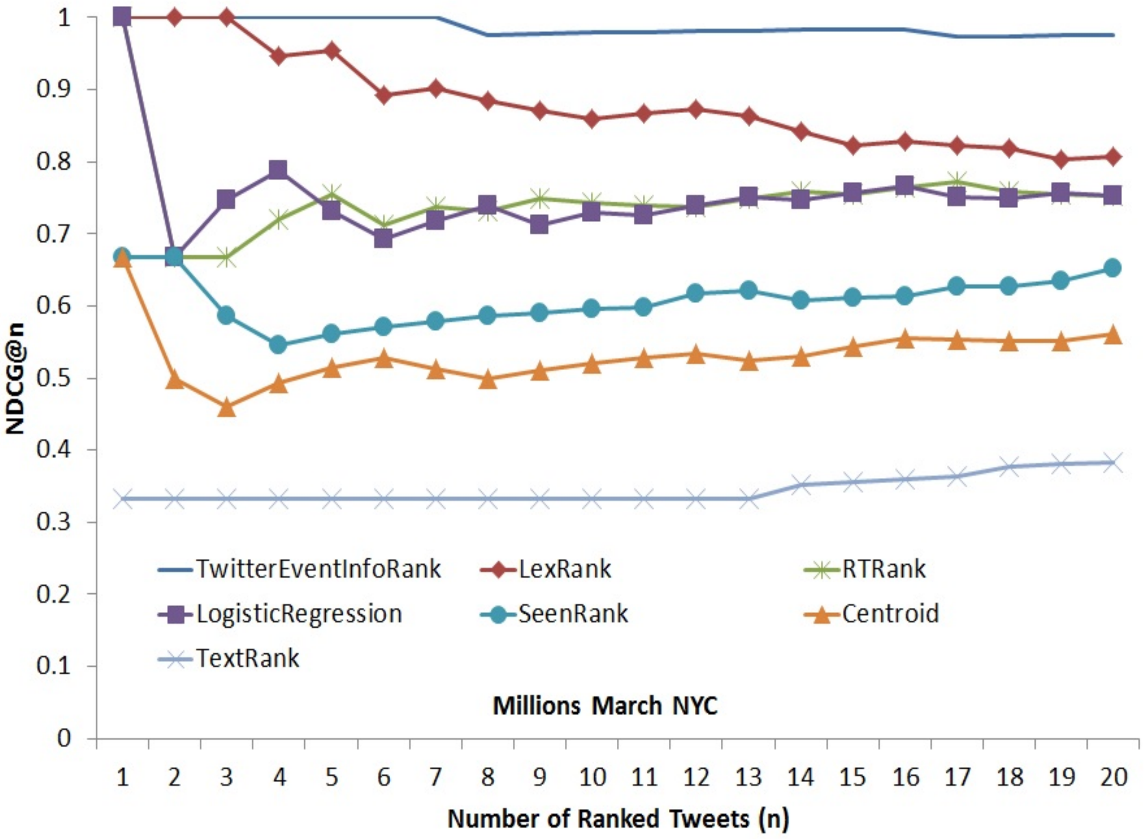


FIGURE 4.10: Performance comparison of ranking techniques using NDCG scores.

TABLE 4.6: 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

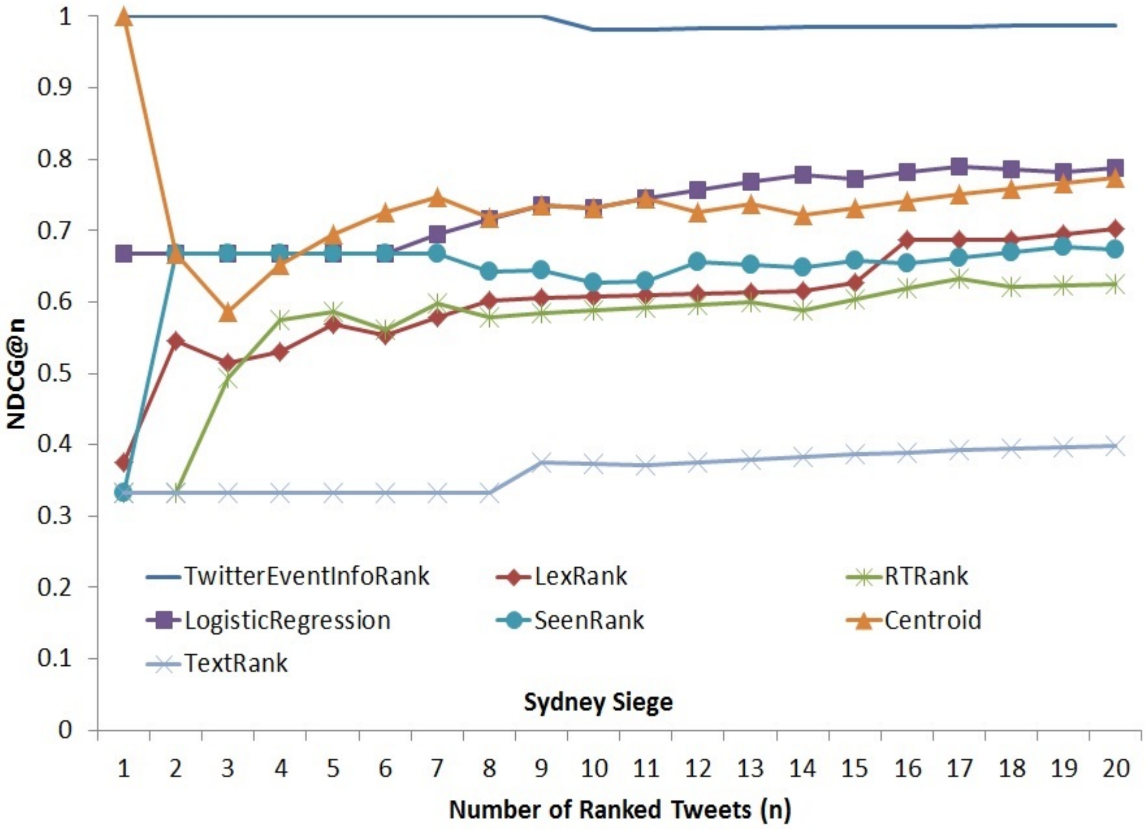
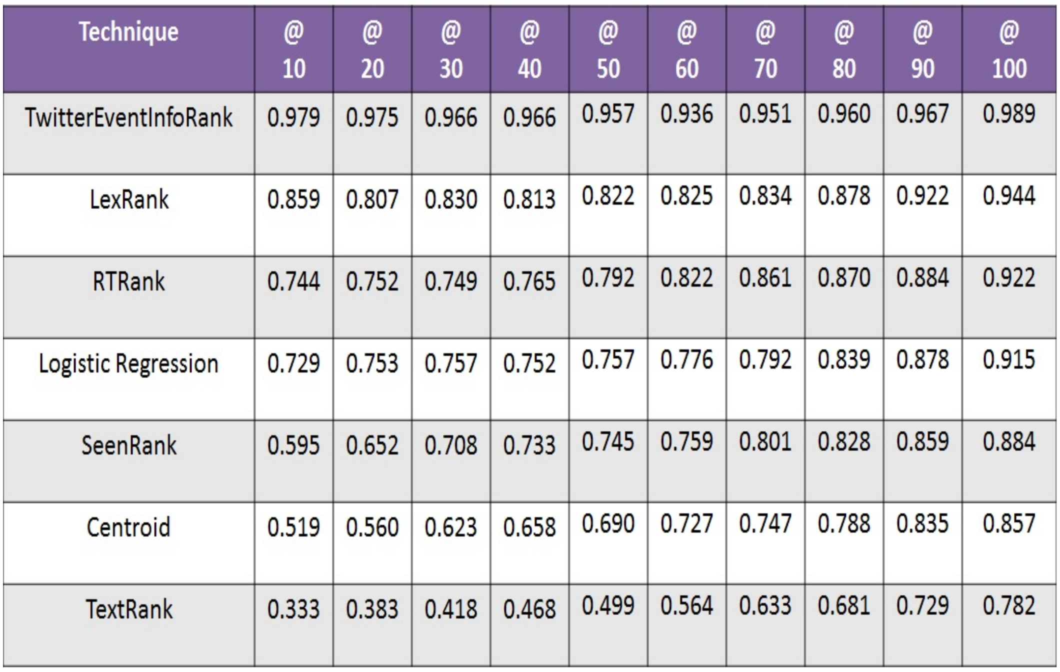
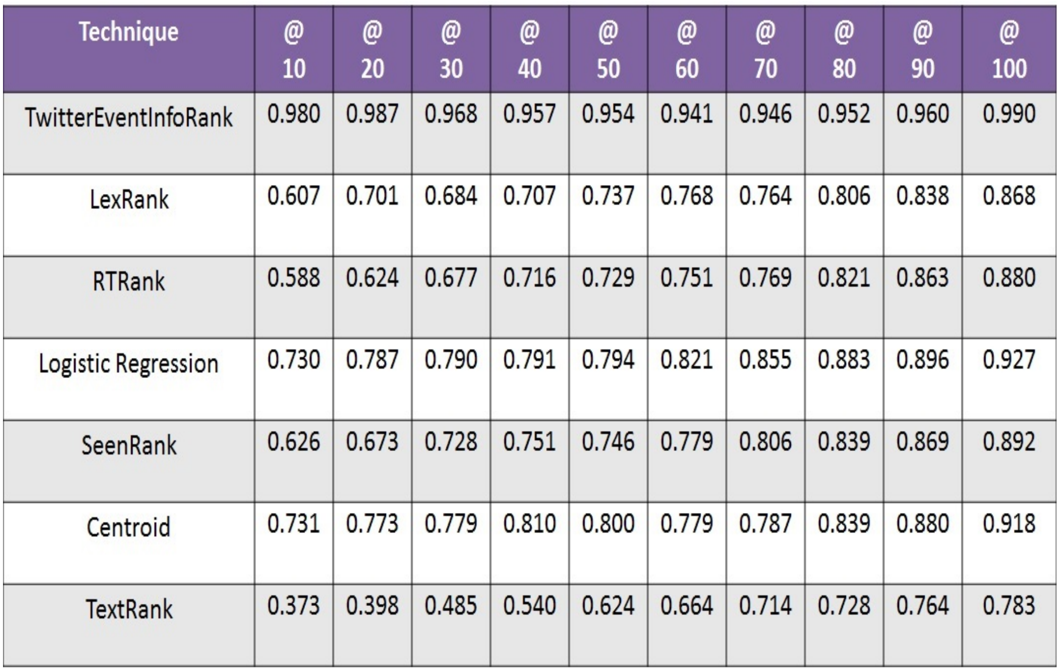


Figure 4.11: Performance comparison of ranking techniques using NDCG scores.



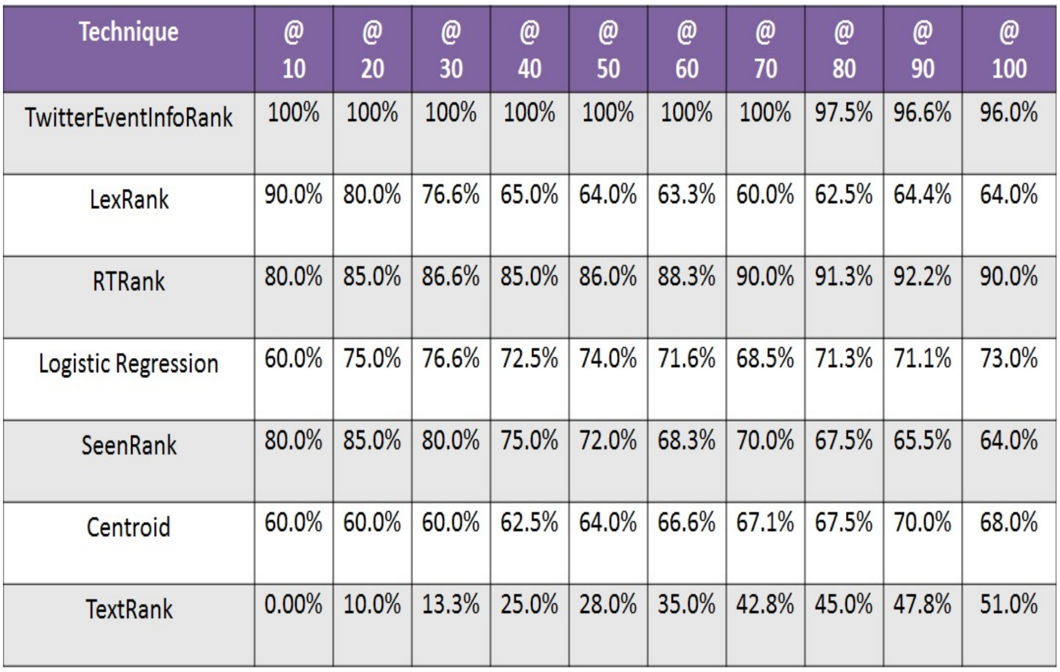
Technique	@ 10	@ 20	@ 30	@ 40	@ 50	@ 60	@ 70	@ 80	@ 90	@ 100
TwitterEventInfoRank	0.979	0.975	0.966	0.966	0.957	0.936	0.951	0.960	0.967	0.989
LexRank	0.859	0.807	0.830	0.813	0.822	0.825	0.834	0.878	0.922	0.944
RTRank	0.744	0.752	0.749	0.765	0.792	0.822	0.861	0.870	0.884	0.922
Logistic Regression	0.729	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 4.12: Performance comparison of ranking techniques using NDCG scores.



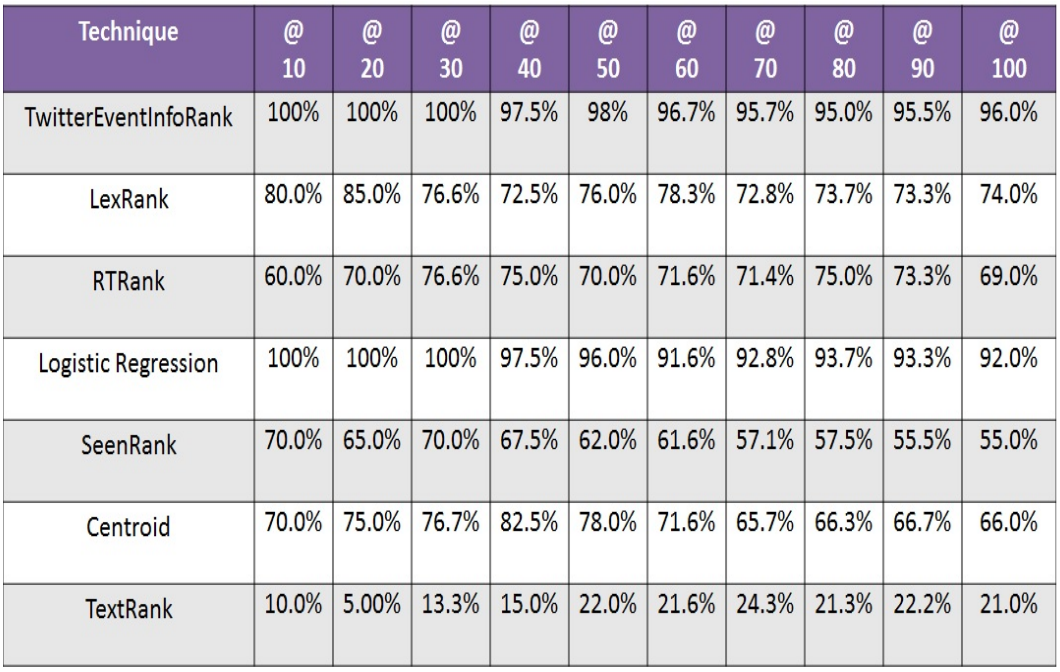
Technique	@ 10	@ 20	@ 30	@ 40	@ 50	@ 60	@ 70	@ 80	@ 90	@ 100
TwitterEventInfoRank	0.980	0.987	0.968	0.957	0.954	0.941	0.946	0.952	0.960	0.990
LexRank	0.607	0.701	0.684	0.707	0.737	0.768	0.764	0.806	0.838	0.868
RTRank	0.588	0.624	0.677	0.716	0.729	0.751	0.769	0.821	0.863	0.880
Logistic Regression	0.730	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 4.13: 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%

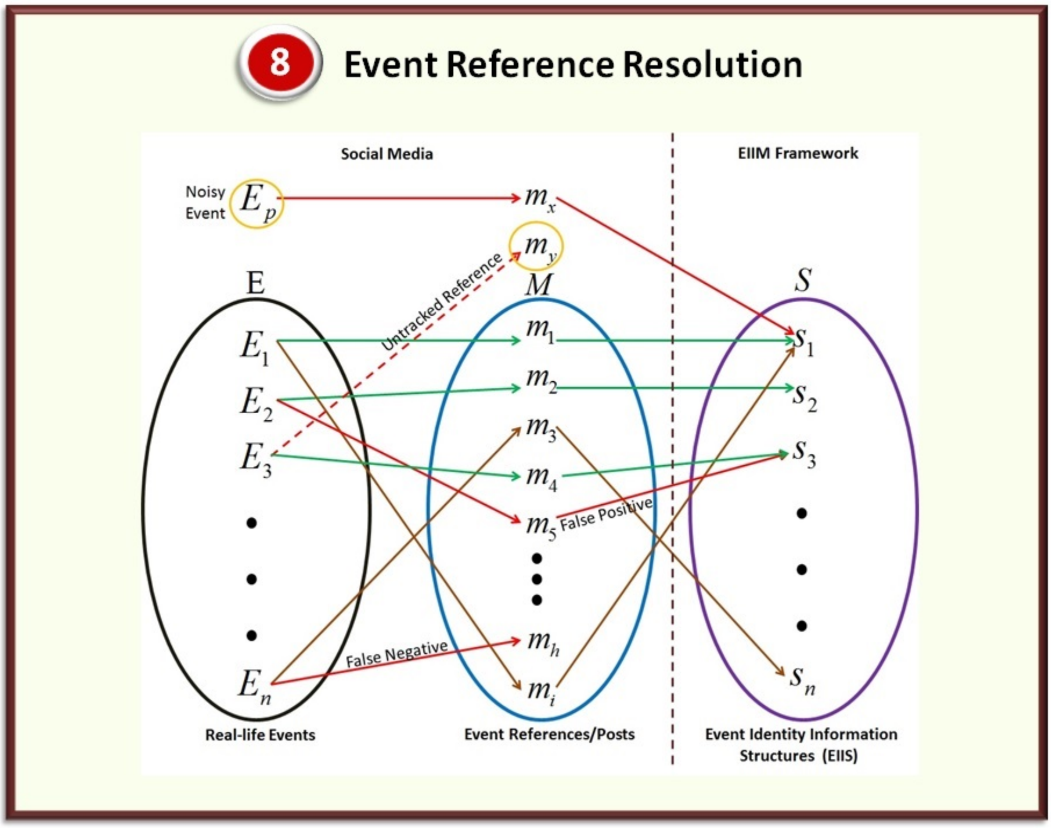
FIGURE 4.14: Performance comparison of ranking techniques using precision scores.



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 4.15: Performance comparison of ranking techniques using precision scores.

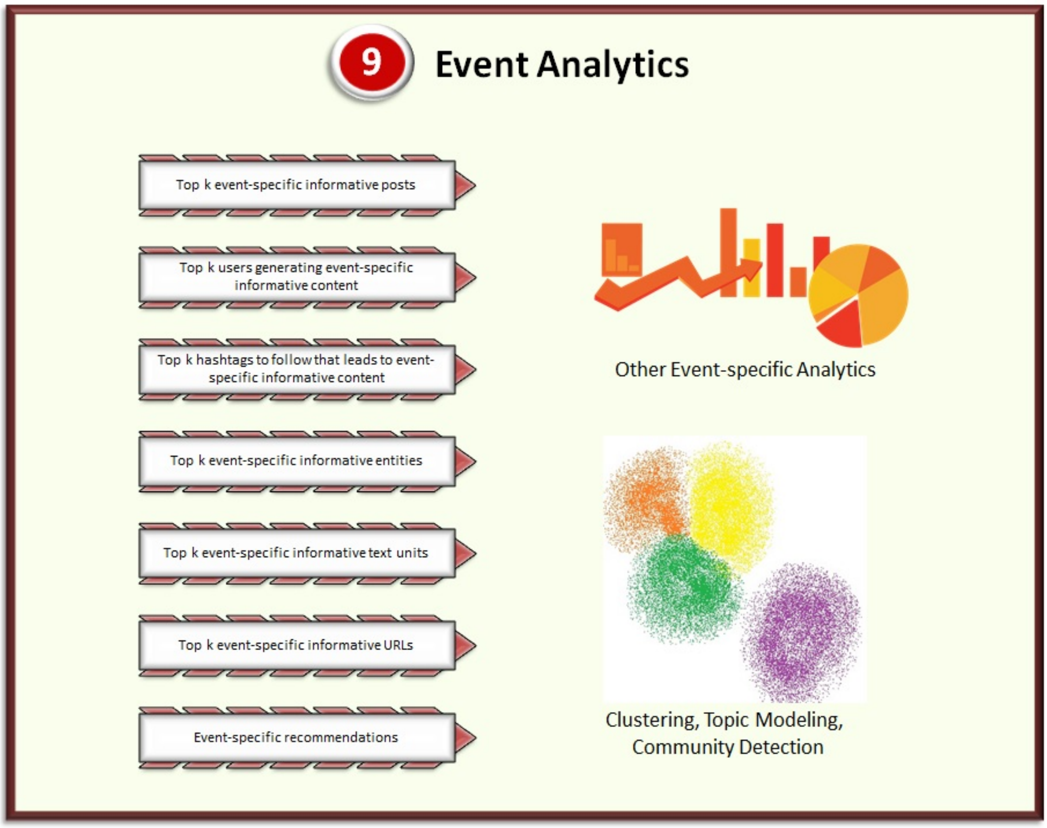
FIGURE 4.16: Event Reference Resolution component of the EIIM life cycle.



4.8 Event Reference Resolution

4.9 Event Analytics

FIGURE 4.17: Event Analytics component of the EIIM life cycle.



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

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...

Chapter 5

Potential Applications of the EIIM Framework

5.1 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' [37] to international sporting events like the 'Winter Olympics' [38] to socio-political [39] and socio-economical [40] events that shook the world such as presidential elections [41], 'Egyptian Revolution' [42], and recessions were covered, analyzed, extrapolated and informed by social media. This prolific event-specific content in social media makes it a promising ground for performing event analytics. Platforms like Geofeedia[1], TwitterStand[2], Twitris[3], Truthy[4], and Tweet-Tracker[5] have developed techniques to provide analytics related to different local and global real-life events.

Monitoring social media has become one of the essential activities of national security agencies for predicting potential threats and mass protests [43]. Social media is being used for tracking terrorism activities [44], collective actions [45], 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/

5.2 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.

5.3 Opinion and Review Mining

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

[7] http://seen.co

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

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

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

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

5.4 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.

5.5 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.

5.6 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 6

Conclusion and Future Work

6.1 Conclusion

6.2 Future Work

6.2.1 Summarizing Event Related Content

6.2.2 Identifying Insightful Opinionated Content Related to Events

6.2.3 Event Topic Modeling

6.2.4 Event-specific Recommendations

6.2.5 Distributed Processing of EventIdentityInfoGraph

6.2.6 Event Ontology for Social Media

Appendix A

Appendix Title Here

Write your Appendix content here.

Bibliography

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

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

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

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

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

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

[7] James Allan. *Topic detection and tracking: event-based information organization*, volume 12. Springer Science & Business Media, 2002.

[8] James Allan, Ron Papka, and Victor Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–45. ACM, 1998.

[9] Giridhar Kumaran and James Allan. Text classification and named entities for new event detection. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 297–304. ACM, 2004.

[10] Yiming Yang, Tom Pierce, and Jaime Carbonell. A study of retrospective and on-line event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 28–36. ACM, 1998.

[11] Kuo Zhang, Juan Zi, and Li Gang Wu. New event detection based on indexing-tree and named entity. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 215–222. ACM, 2007.

[12] Vasileios Hatzivassiloglou and Elena Filatova. Domain-independent detection, extraction, and labeling of atomic events. 2003.

[13] Vasileios Hatzivassiloglou, Luis Gravano, and Ankineedu Maganti. An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 224–231. ACM, 2000.

[14] Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi. Simple semantics in topic detection and tracking. *Information Retrieval*, 7(3-4):347–368, 2004.

[15] David Ahn. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8. Association for Computational Linguistics, 2006.

[16] Heng Ji and Ralph Grishman. Refining event extraction through cross-document inference. In *ACL*, pages 254–262, 2008.

[17] Elena Filatova, Vasileios Hatzivassiloglou, and Kathleen McKeown. Automatic creation of domain templates. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 207–214. Association for Computational Linguistics, 2006.

[18] Jianshu Weng and Bu-Sung Lee. Event detection in twitter. *ICWSM*, 11:401–408, 2011.

[19] Jagan Sankaranarayanan, Hanan Samet, Benjamin E Teitler, Michael D Lieberman, and Jon Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, pages 42–51. ACM, 2009.

[20] Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics, 2010.

[21] Tye Rattenbury, Nathaniel Good, and Mor Naaman. 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.

[22] Ling Chen and Abhishek Roy. Event detection from flickr data through wavelet-based spatial analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 523–532. ACM, 2009.

[23] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[24] Sarita Yardi and Danah Boyd. Tweeting from the town square: Measuring geographic local networks. In *ICWSM*, 2010.

[25] Edward Benson, Aria Haghighi, and Regina Barzilay. Event discovery in social media feeds. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 389–398. Association for Computational Linguistics, 2011.

[26] Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. Linking online news and social media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 565–574. ACM, 2011.

[27] Rinkesh Nagmoti, Ankur Teredesai, and Martine De Cock. Ranking approaches for microblog search. In *2010 IEEE/WIC/ACM International joint conference on Web Intelligence-Intelligent Agent Technology (WI-IAT)*, volume 1, pages 153–157. IEEE Computer Society, 2010.

[28] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[29] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.

[30] D Tunkelang. A twitter analog to pagerank. *The Noisy Channel*, 2009.

[31] V Hallberg, A Hjalmarsson, J Puigcerver, C Rydberg, and J Stjernberg. An adaptation of the pagerank algorithm to twitter world. 2012.

[32] Richard McCreadie and Craig Macdonald. Relevance in microblogs: Enhancing tweet retrieval using hyperlinked documents. In *Proceedings of the 10th conference on open research areas in information retrieval*, pages 189–196. LE CENTRE

DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2013.

[33] Jan Vosecky, Kenneth Wai-Ting Leung, and Wilfred Ng. Searching for quality microblog posts: Filtering and ranking based on content analysis and implicit links. In *Database Systems for Advanced Applications*, pages 397–413. Springer, 2012.

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

[35] 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.

[36] 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.

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

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

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

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

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

[42] 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.

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

[44] 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.

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