University of Arkansas at Little Rock

Doctoral Dissertation

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 dissertation 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 such as Facebook, Twitter and Google Plus, there has been voluminous growth in the digital footprints of real-life events on the Internet. The user-generated colloquial and concise textual content related to different types of real-life events, available in these websites, acts as an extremely useful source for researchers and organizations for extracting valuable and insightful information. There has been significant improvement in natural language processing techniques for mining formal and long textual content commonly found in newspapers. It is still a challenging task to mine textual information from the social media channels producing terse, informal and noisy text with an unusual grammatical structure. For a real-life event of interest it is necessary to detect and store informative event-specific signals from the noisy social media channels that allows to distinctly identify the event among all others, and characterizes it for extracting actionable insights. These event-specific cues also form 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, data journalism, cyber security, event management, among others. Thus, there is a need of a generic framework that can collect the textual content related to a real-life event, extract event-specific information from it and persistently maintain the information for tracking newly produced content as the event evolves, and provide updated event analytics. The patent-pending work presented in this dissertation establishes the design and implementation of an extendable framework that enables collection, extraction and persistent management of 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 critical component of the EIIM life 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. 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 South Asian 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 would also like to thank Dr. Nitin Agarwal, who supported me in the initial days of my PhD.

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 in social media, from 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.

Chapter 1

Dissertation Overview

Social media is a paradigm shift in the way people communicate with each other. It has grown from being just a medium to a global medium of communication between people. Different types of social media platforms provide multiple venues for people to share firsthand experiences and exchange information about real-life events. It has become an indispensable means for disseminating news and real-time information about current events, using websites such as Twitter, Facebook, Instagram, Flickr, Youtube, Google Plus and Vine. These applications allow users to post short textual messages accompanied by images and videos. At the same time users also share their detailed journalistic experiences in the form of diaries through blogging platforms such as Blogger, Wordpress and Medium. Studies have shown the importance of social media platforms as a news circulation service [2], and a source for gauging public interest and opinions [3–6]. Its efficacy as a real-time, citizen-journalistic source of information has been recently harnessed in the detection, extraction and analysis of real-life events [7–9]. The activities of users producing content in social media has also been studied for gaining deep insights about how users form communities around topics related to real-life events [10–12] that lead to collective action [13, 14].

With the popularity of social media there has been proliferation of unstructured textual content on the Internet about different real-life events. Tracking social media for useful content such as live reporting of an event and recent updates can be harnessed to identify important nuggets of information. It can lead to identification and analysis of insightful user-generated information about named entities (people, place, organization, etc). Event summaries can be generated from the identified event-specific informative content. Actionable event-specific insights can be extracted such as, "what is happening", "who is involved", "where is it happening", "when did it happen", and so on. There are tremendous applications in the areas of real-life event analysis, data journalism, event management, opinion mining, online targeted marketing, cyber security, among others. Thus, there is a need of a generic framework that has the following capabilities:

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

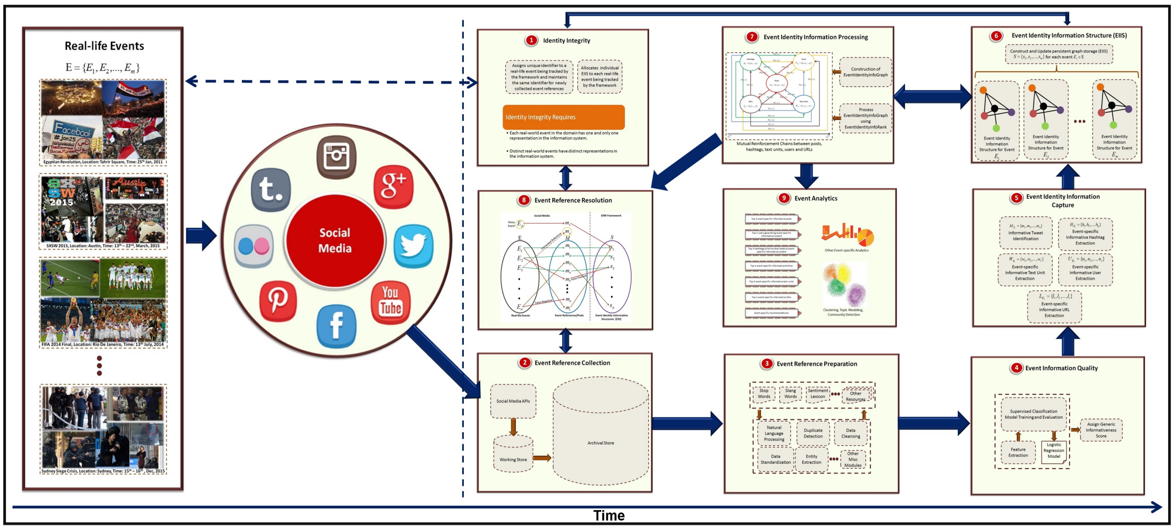
The problem of collecting and extracting event identity information from social media is very similar to the task of event detection and tracking from newswires [15, 16]. However, in this dissertation, we add new components that creates persistent identity structures of an event and update it with new information over time. In order to make our task well-defined we focus on the problem of tracking a pre-specified set of events. Also, the domain of social media poses additional challenges. News articles most often adhere to grammatical, syntactical and formal structures of writing, that are not common in the realm of social media. The user-generated content in social media is most often colloquial, short, noisy, and lacks proper grammatical structure. This makes it a challenging task for even the state-of-the-art natural language processing techniques to extract useful information and to perform tasks like entity extraction and parts-of-speech tagging that lies at the core of the previous research on event detection and tracking.

The work presented in this dissertation establishes the conceptual design and implementation of a framework capable of collecting, extracting and persistently managing event identity information from user-generated short textual content shared in social media (shown in Figure 1.1). The approach of the presented work is from the perspective of Entity Identity Information Management (EIIM) [1], with basic tenets of information quality at its core. Towards this objective, different challenges of mining high quality information from social media text is discussed and a patent-pending, novel approach to the challenges of identifying event-specific informative content is explained. This approach is a critical component of the framework. The dissertation further explores the applications of the research and concludes by discussing future directions of the work.

Some of the main contributions of the work are:

- Extending the Entity Identity Information Management (EIIM) model [1] from the closed world domain of Master Data Management (MDM) to the open and unstructured domain of social media.

FIGURE 1.1: Event Identity Information Management (EIIM) Life Cycle for user generated textual content in social media



- The design and implementation of an *Event Identity Information Management* framework that is capable of tracking and identifying event-specific information from short user-generated textual content in social media. Towards this objective a data processing pipeline named *Event Identity Information Management Life Cycle* is developed, which is capable of :
  - Collecting event related real-time content generated in social media.
  - Pre-processing them using natural language processing techniques.
  - Identifying high-quality sources of information.
  - Extracting event-specific information in order to create *Event Identity Information Structures* (EIIS) for persistently storing and characterizing the salient and high-quality event related information.
  - Identifying event-specific informative content produced in social media.
- Implementation of a supervised classifier in the domain of short and informal social media textual content, for segregating high-quality informative messages having higher chances of containing event related information from the low-quality non-informative ones.
- Analysis of informative and non-informative event related content from more than 3.8 million short textual social media messages.
- A novel model based on the principle of mutual reinforcement that takes into account the semantics of relationships between short textual *social media messages*, *hashtags*, *text units*, *URLs* and *users*, and represent them in a graph structure - *EvenIdentitytInfoGraph*. A scalable graph processing iterative algorithm

-*EventIdentityInfoRank*, is implemented for ranking the nodes of the *EventIdentityInfoGraph*. The algorithm is capable of simultaneously ranking *social media messages*, *hashtags*, *text units*, *URLs* and *users* in terms of event-specific informativeness providing deeper insights into the identity of an event.

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

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

Related Filed Patent

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

Related Award

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

Related Publications

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

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

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

- Nitin Agarwal, **Debanjan Mahata**, and Huan Liu. *Time-and Event-Driven Modeling of Blogger Influence.* Encyclopedia of Social Network Analysis and Mining. Springer New York, 2014. 2154-2165.
- **Debanjan Mahata** and Nitin Agarwal. *Learning from the crowd: An Evolutionary Mutual Reinforcement Model for Analyzing Events.* Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. IEEE, 2013.
- Nitin Agarwal, and **Debanjan Mahata**. *Grouping the Similar among the Disconnected Bloggers.* Social Media Mining and Social Network Analysis: Emerging Research (2013), 54.
- **Debanjan Mahata**, and Nitin Agarwal. *What does everybody know? identifying event-specific sources from social media.* IEEE Fourth International Conference on Computational Aspects of Social Networks (CASoN), 2012.
- **Debanjan Mahata** and Nitin Agarwal. *Analyzing Event-specific Socio-Technical Behaviors Through the Lens of Social Media.* The International Sunbelt Social Network Conference (Sunbelt XXXII) organized by the International Network for Social Network Analysis (INSNA), March 12-18, 2012, Redondo Beach, California.
- Vivek Kumar Singh, **Debanjan Mahata**, and Rakesh Adhikari. *Mining the blogosphere from a socio-political perspective.* IEEE International Conference on Computer Information Systems and Industrial Management Applications (CISIM), 2010.
- Vivek Kumar Singh, Rakesh Adhikari, and **Debanjan Mahata**. *A clustering and opinion mining approach to socio-political analysis of the blogosphere.* IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2010.

Related Submitted Publications

- **Debanjan Mahata**, John R. Talburt, Vivek Kumar Singh and Rajesh Piryani; *Chatter that Matter: A Framework for Identifying and Ranking Event-specific Informative Tweets.* 18^{th} International Conference on Text, Speech and Dialogue, Plzen, Czech Republic (Notification Due: May 10, 2015)
- **Debanjan Mahata**, John R. Talburt and Vivek Kumar Singh; *A Framework for Collecting, Extracting and Managing Event Identity Information from Twitter.* 20^{th} International Conference on Information Quality, M.I.T, Boston (Notification Due: April 30, 2015)

- **Debanjan Mahata**, John R. Talburt and Vivek Kumar Singh; *From Chirps to Whistles : Discovering Event-specific Informative Content from Twitter*. Proceedings of the 7^{th} Annual ACM Web Science Conference. ACM, 2015, Oxford, England (Notification Due: April 30, 2015)

The rest of the thesis is organized as follows:

Chapter 2 gives an overview of the different social media websites and their common characteristics. It also looks at the different perspectives of defining an event and gives the definition of events in social media as accepted by the presented work. Finally, it defines the problem of Event Identity Information Management from Social Media whose solution and application is extensively discussed throughout the rest of the dissertation.

Chapter 3 reviews the existing literature related to the topic of the dissertation and highlights the challenges in applying previously available techniques to the domain of social media. It also discusses the similarities and dissimilarities of our work with the previous ones, and identifies the areas of our novel contributions that makes it different from the available techniques.

Chapter 4 gives a detailed explanation of the challenges that are relevant to the problem of this dissertation. It discusses different scenarios in which these problems occur and point to their solutions as proposed in this dissertation.

Chapter 5 presents a detailed discussion of the *Event Identity Information Management Life Cycle*, that is proposed as a solution to the problem that is posed in this dissertation. It goes through all the components of the life cycle and gives a detailed explanation of the design choices, implementation and their working.

Chapter 6 discusses about the baselines selected for evaluating the approaches presented in the dissertation and compares their performance with them.

Chapter 7 highlights the potential real-life application of the *Event Identity Information Management* framework implemented in this thesis.

Chapter 8 draws conclusions of the work presented in this thesis and points to future directions of the work.

Chapter 2

Social Media and Real-life Events

2.1 Social Media

Social media is defined as a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content [17]. These Internet-based applications broadly ranges from blogs, microblogs, media sharing webistes, social bookmarking websites, social news to social networking websites. Brief description of the most popular types of social media websites is given below.

2.1.1 Blogs

A blog can be defined as a website that displays, in a reverse chronological order, the entries by one or more individuals and usually has links to comments on specific postings. Blogs often provide opinions, commentaries, or news on a particular subject, such as food, politics, or local news. Some of them also function as personal online diaries. Most of the time the entries of a blog is archived and is accessible at a later time. For the purpose of constant syndication, RSS or XML feeds for the blogs are made available. An individual entry in a blog is known as a blog post. A typical blog post can combine text, images and links to other blogs, web pages and other media related to its topic. The universe of all the blogs on the Internet is known as blogosphere [11].

2.1.2 Microblogs

Microblogs are similar to blogs, but a shorter version of it. Most of the microblogging websites pose limitations on the length of an individual post. Twitter, one of the most popular microblogging website has a limitation of 140 characters. This makes the textual posts in these platforms, extremely concise. Users often associate URLs that lead to external sources of information related to the posts. A post may also contain attached image or video. The microblogging services mostly focus on short updates that are pushed out to anyone subscribed to receive the updates. This is made possible by enabling the users to form directed networks of *friends* and *followers*. The *followers* of an user are entitled to get all the updates posted by him. Mostly these updates are public.

2.1.3 Media Sharing

Media sharing services allow its users to upload and share various multimedia content such as pictures and videos. Most services have additional social features such as profiles, commenting, etc. The most popular are Instagram, Pinterest, YouTube and Flickr. The media elements are often enriched with geographical and topical “tags” by the users who create them and the consumers who browse them. These tags acts as very useful meta data and allow automated programs to leverage them for efficient organization and retrieval of the videos and images that otherwise have very less textual content.

2.1.4 Social Bookmarking

These are the genre of social media services that allow its users to save, organize and manage links to various websites and resources around the internet. Most allow to “tag” URLs for making them easy to search and share. The most popular are Delicious and StumbleUpon. Some of the services like StumbleUpon also allow their users to form friendship networks. These websites often provide different browsing experiences through interfaces that help the users to search for most recent tags, most popular ones, and so on.

2.1.5 Social News

Social news websites allow people to post various news items or links to articles that are external to the website, and then allows its users to cast their “vote’ on the items. The voting is the core social aspect as the items that get the most votes are displayed most prominently. This makes it an ideal crowdsourced news platform. It is up to the community of users to decide which news items gets seen by more people. Users can also “tag” the news stories and comment on them. The most popular are Digg and Reddit.

2.1.6 Social Networking

Social networking websites are the ones that allow its users to connect with each other and form networks. The connections are generally non-directional and reciprocal. Two users who are connected to each other are considered as *friends*. Usually the users in these webistes have a profile that presents the personal information of the user as provided by him. The users have various ways to interact with other users, and also sometimes have the ability to set up groups. These social networks may be based on a certain theme such as interests, location, and profession. Facebook is the most popular personal social network and LinkedIn is the most popular professional network.

Some of the other types of websites that can also be categorized as social media services are, social messaging services, collaboration tools, rating or review sites, personal broadcasting tools, virtual worlds, and group buying. Table 2.1, lists popular social media websites in different categories. Some of the websites may overlap and fall into multiple categories due to the broad range of services provided by them. For example, Facebook is not only a popular social networking website, but also a widely used social messaging service.

TABLE 2.1: Popular social media websites belonging to different categories.

Category	Popular Social Media Websites
Blogs	Blogger, Medium, Wordpress, Squarespace
Microblogs	Twitter, Tumblr, Posterous
Media Sharing	Flickr, Instagram, YouTube, Vimeo, Dailymotion, Metacafe, Viddler, Pinterest
Social Bookmarking	Delicious, StumbleUpon, Scoop, Slashdot
Social News	Digg, Reddit, Newsvine, Propeller
Social Networking	Facebook, Google Plus, LinkedIn, Ello, CafeMom, Gather, Fitsugar
Virtual Worlds	Second Life, World of Warcraft, Farmville
Group Buying	Groupon, Living Social, Crowdsavings
Personal Broadcasting	Blog Talk radio, Ustream, Livestream
Review/Rating	Amazon ratings, Angie's List
Collaboration Tools	Wikipedia, WikiTravel, WikiBooks
Social Messaging	WhatsApp, Viber

According to Pew Research Center Facebook, LinkedIn, Pinterest, Instagram and Twitter are the top five most popular social media websites used by American adult Internet users[1]. The number of active users world-wide for all the five social media sites is shown in Table 2.2. The numbers are obtained from the official pages of the respective websites.

[1]http://www.pewinternet.org/2015/01/09/social-media-update-2014/

Table 2.2: Number of active users for the top five social media websites used by American adults.

Social Media Website	Number of Active Users
Facebook	1.31 billion
LinkedIn	347 million
Pinterest	70 million
Instagram	100 million
Twitter	289 million

2.2 Characteristics of Social Media Websites

All the above social media websites exhibit certain common characteristics that is also responsible for their wide usage and huge popularity. We revisit some of the characteristics as already suggested by Agarwal et al. [18] in the context of this dissertation.

1. **Accessibility:** Social media websites are freely available to whoever has an Internet connection. This makes these websites easily accessible all over the world. One of the latest initiatives by Facebook and Google is to make social media accessible even to the most remote corner of the world through their Internet.org[2] and "Loon for All"[3] projects, respectively. With the popularity of hand held devices and increase in the Internet bandwidth, social media is accessible to anyone who has a smart phone and can use it. This is unlike the mainstream media or the print media, to which people subscribe and buy in the form of magazines, newspapers, journals, etc. Also, the mainstream media can be easily controlled by the government that may lead to propagation of biased information. For example, during the "Egyptian Revolution of 2011", the mainstream media was biased, regulated by the government, and did not portray the true picture of the situation in Egypt. On the other hand, it was social media through which people discussed about the actual atrocities of the government and grouped together to incite the entire revolution [19].

2. **Permanence:** Social media websites show dynamic nature and the content can be altered any time. Users can easily edit the content shared by them. On the other hand the traditional print media and television media is not at all dynamic. Once an article is printed in a magazine/newspaper, or a television show is recorded and broadcasted, it cannot be changed.

3. **Reach:** As already presented in Table 2.2, some of the popular social media websites have a reach of billions. Moreover, the ability of an individual user to

[2]http://internet.org/
[3]http://www.google.com/loon/

simultaneously network with many other users makes this reach more effective than any other means of communication. These connections acts as networks of information flow, which helps in spreading any kind of information at a lightning speed [20]. Also it provides equal opportunity to everyone for reaching their intended audience, unlike the traditional mainstream media. This characteristic is now regularly used by politicians for launching election campaigns and reaching out to people in social media [21]. The marketers also leverage social media to a great extent. Event managers also take advantage of it, due to which social media has become an integral part of event management for getting connected with the event audience [22]. Social media is regularly used during planned events for making announcements, building and tracking audience, building focused communities, developing public relations [23], and targeted marketing [24]. Due to easy reachability in social media, a focused group of people can also get together very quickly and organize events such as flashmobs and protest movements [12].

4. **Recency:** The time lag at which communication can take place and information can flow is almost zero for the social media websites. Content is produced and communicated in real-time. Once this content is consumed, the users discuss about it instantaneously. Due to the reach and recency of social media it can make people aware of newsworthy events at a faster pace than traditional maintream media [2, 25]. It might take hours, days, or sometimes months to present news or an event through mainstream media like newspapers, magazines and television. For example, the death of Bin Laden and the entire covert opertation was reported in Twitter even before the US president made an official announcement in the mainstream media [26]. The users in social media were not only aware of the event but were also sharing and discussing it with great zeal. Another example is of the theater shootings in Colorado [27]. The shooting incident was reported, covered and analyzed in real-time, with traditional news media lagging behind social media by several minutes.

5. **Usability:** Social media sites are extremely easy to use and are user-friendly. An user does not require special training or skills to create content in a social media website. Whoever, can type or use a device connected to the Internet that enables typing of text can share information in social media. Therefore, the operational cost of any social media application is mostly negligible. This is not the case for mainstream media. In order to report an event one needs to be skilled and specially trained. Also, the printing and telecasting of any event has to go through many other processes and has to be finally approved by the editor. This makes traditional media unusable by common people. The use of social media at the time of Internet blackout during the "Egyptian Revolution of 2011" is a great example of

its usability [28]. The Egyptian government had throttled the Internet connection and there was an Internet blackout for hours in order to stop the spreading of messages in social media, which was the major media of communication that led to the protest movements and finally to the revolution. In order to counter attack the government and to allow the Egyptians to use social media for giving latest updates, Google and Twitter launched a service that enabled them to leave a voicemail on a specific number. This voicemail was then posted in Twitter as a text message. Thus, people who didn't have a smart phone or didn't know how to use one could also post messages in social media.

2.3 Defining Events from Different Perspectives

2.4 Events in Social Media

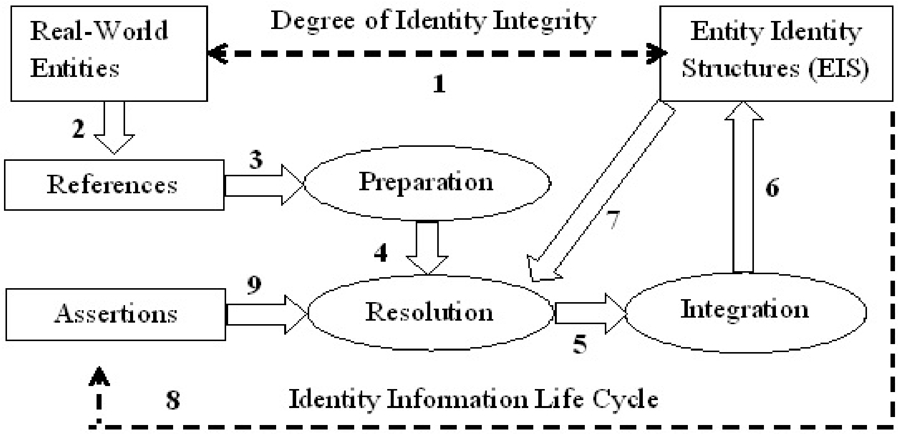
2.5 Background: Entity Identity Information Management (EIIM) in Master Data Management

The idea of Entity Identity Information Management (EIIM) as defined by Zhou et al. [1], is the collection and management of identity information of real-world entities with the goal of sustaining *entity identity integrity*. *Entity identity integrity* is one of the basic tenets of data quality that applies to the representation of a given domain of real-world entities in an information system [29]. In order to maintain the property of *entity identity integrity* following conditions should be satisfied:

1. Each real-world entity in the domain has one and only one representation in the information system.
2. Distinct real-world entities have distinct representations in the information system.

Their model of EIIM was motivated by the problem of entity resolution in information systems, particularly in the domain of MDM (Master Data Management). They define entity resolution as the process of determining whether two references to real-world objects in an information system are referring to the same object, or to different object [29]. The EIIM life cycle as proposed by them is an iterative process that combines entity resolution and data structures representing entity identity into specific operational configurations (EIIM configurations, as shown in Figure 2.1), that when executed in concert, work to maintain the entity identity integrity of master data over time. The EIIM framework is implemented by developing open source software known as OYSTER[4].

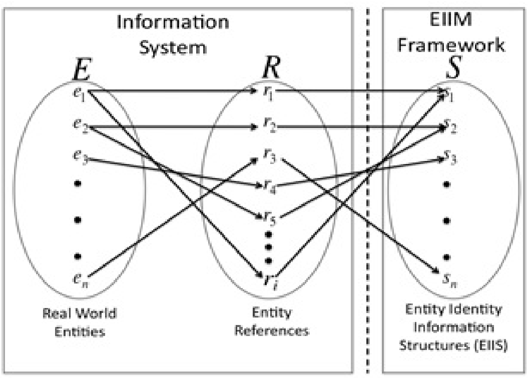
Figure 2.1: EIIM components and their interactions as proposed in [1]



Some of the definitions as specified by the current EIIM model are:

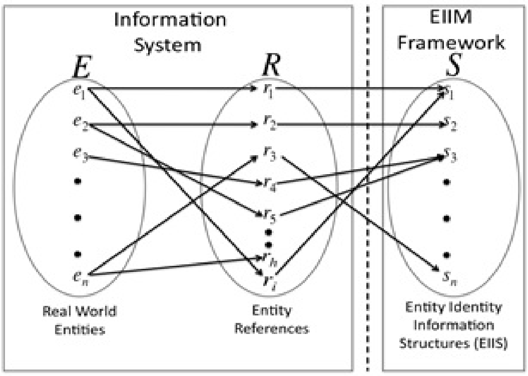
- **Definition 1.** *An **entity** (e_i) is defined as a real-life object that has a distinct identity.*
- **Definition 2** ***Entity Identity Information*** *is defined as a set of attributes of a given entity that distinctly characterizes it and allows that entity to be distinguished from all the other entities maintained by the framework.*
- **Definition 3.** *An **Entity Identity Information Structure (EIIS)** (s_i), is defined as a data structure that can persistently and efficiently store, retrieve, and manipulate entity identity information.*

Figure 2.2: Entity Identity Integrity in EIIM process.



[4] http://sourceforge.net/projects/oysterer/

Figure 2.3: Misjudgments made by EIIM process.



Therefore, ideally in an information system, if $E = \{e_1, e_2, ..., e_n\}$ represents a finite set of entities, $R = \{r_1, r_2, ..., r_m\}$ represents a finite set of references to the entities, and $S = \{s_1, s_2, ..., s_n\}$ represents a finite set of EIIS maintaining identity information of the entities then there should be one-to-one correspondence between the real-life entities ($\in E$) and the EIIS ($\in S$) representing their identity information. Also, the references ($\in R$) of a particular entity ($\in E$) should always map to one and only one EIIS ($\in S$) maintaining its identity information. This is shown in Figure 2.2. Such a situation ensures that the condition of entity identity integrity is satisfied by the information system. One of the main aims of EIIM is to satisfy the conditions of entity identity integrity along with persistently maintaining the entity identity information.

The current EIIM model deals with a closed environment of an information system where there is fixed number of entities along with fixed number of references to them. In an ideal situation the EIIM process should always satisfy the conditions of entity identity integrity as shown in Figure 2.2, and previously explained. However, in practice, all the references to an entity in the information system might not get mapped to the EIIS maintained for that particular entity due to misjudgments made by the automated processes as shown in Figure 2.3. This might result in *false negative* and *false positive* errors. A *false negative* error arises when the system fails to map a reference of an entity to its corresponding EIIS. This is shown in Figure 2.3, where the system fails to map the reference r_h ($\in R$) of entity e_n ($\in E$) to an EIIS ($\in S$). A *false positive* error arises when the system maps two references of different entities to a single EIIS. This is shown in Figure 5, where the system wrongly maps reference r_5 ($\in R$) of entity e_2 ($\in E$), to the EIIS s_3 ($\in S$) being maintained for entity e_3 ($\in E$). Such a situation creates dissonance between the actual identity of the real-world entities being stored in the information system and their identities interpreted by the automated processes, resulting in low entity identity integrity of the system. Asserted resolutions are introduced in order to deal with such problems (shown in Figure 2.1.).

The EIIM processes and life cycle is a step ahead of the basic record linking process that identifies references to same entities for a given dataset. The goal of EIIM is to consistently label references to the same entity with the same identifier across different datasets processed at different times. Through the management of persistent entity identity structures, EIIM provides an added functionality for an entity resolution system to create and assign persistent entity identifiers that do not change from process to process. The current EIIM can also be thought of as forming a nexus between ER and MDM by adding an explicit longitudinal dimension to the management of identity information. The EIIM model proposed in the presented research expands the current model into the unstructured domain of social media, bringing in new challenges and devising new techniques for solving them. The next section gives a detailed discussion and definition of the problem of extending the EIIM model to social media.

2.6 Problem of Event Identity Information Management (EIIM) in Social Media

FIGURE 2.4: Event identity information for real-life event, 'Egyptian Revolution 2011'.

The definitions as given in the previous section also hold true for the work presented in this dissertation. The only differences are:

1. We consider a **real-life event** E_i as an individual **entity**, and assume that every **real-life event** also has a distinct identity. Only a pre-specified set of events $\Xi = \{E_1, E_2, ..., E_n\}$ are considered, with their ordered stream of references $M = \{m_{1_{E_1}}, m_{2_{E_1}}, ..., m_{i_{E_i}}, ..., m_{n_{E_n}}\}$ collected from social media instead of a fixed set of references already present in a closed information system.

2. Instead of **Entity Identity Information**, we define **Event Identity Information** as a set of attributes that distinctly characterizes it and allows that event to be distinguished from all other events maintained by the framework. For example, names of the main active participants (people, organizations, etc), popular keywords, timespan of occurrence and place of occurrence could be identity information for a real-life event as shown in Figure 2.4.

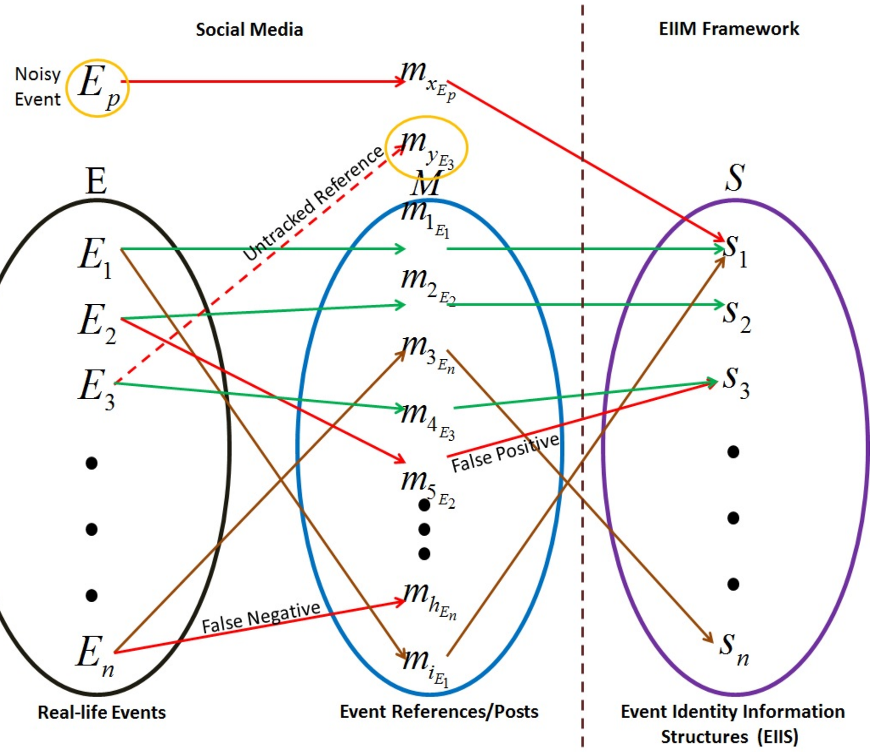
3. Instead of **Entity Identity Information Structure (EIIS)**, we define **Event Identity Information Structures (EIIS)**, for real-life events as a data structure that can persistently and efficiently store, retrieve, and manipulate entity identity information. A set of EIIS, $S = \{s_{E_1}, s_{E_2}, ..., s_{E_n}\}$ is maintained, corresponding to each event $E_i \in \Xi$

4. Lastly, this dissertation solves the problem of *Event Identity Information Management* in the open and unstructured domain of social media, that gives rise to new challenges. It added complexity to the original problem of EIIM, for which following steps are taken (explained in Chapter 5):

   - A completely new life cycle of data processing components had to be introduced that is more suitable for processing unstructured references of events in social media.
   - What consists of **Event Identity Information** became losely defined due to the unstructured and highly dynamic nature of the environment. This led to development of new techniques that can identify identity information of an event w.r.t time, extract it from the references and store them in the **Event Identity Information Structures (EIIS)**.
   - A complete new representation of **Event Identity Information Structures (EIIS)** is proposed, which is more suitable for managing the identity information of events in a dynamic and unstructured environment.

Events have been defined from various perspectives and in different contexts as already discussed in the previous section. In the context of the work presented in this dissertation we adopt a definition similar to [30].

FIGURE 2.5: Relation between elements of Ξ, M and S in Event Identity Information Management from social media.



Event: *An event is defined as a real-world occurrence* (E_i) *with an associated time period* T_{E_i} $(t_{E_i}^{start}\text{-}t_{E_i}^{end})$ *and a time ordered stream of social media references* $M_{E_i} = \{m_{1_{E_i}}, m_{2_{E_i}}, ..., m_{n_{E_i}}, ..., m_{z_{E_i}}\}$*, of substantial volume, discussing about the event and posted in time* T_{E_i}.

Users in social media post about real-life events in huge volumes and with high velocity. This results in multiple footprints of an event in different social media channels. The footprints could be in the form of textual posts, images, videos or other types of multimedia documents. For example, following are three different tweets retrieved for the same event "Sydney Siege Crisis":

1. RT @cnni: Hostage taker in Sydney cafe demands ISIS flag and call with Australian PM, Sky News reports. http://t.co/a2vgrn30Xh #sydneysiege
2. "@carlyeinfeld: Whats with the barking dogs in the background of the briefing on the #sydneysiege right now?!" Police dogs.
3. @Channel4News The least you could do is to pixelate the faces in the window #sydneysiege

We consider these footprints as references to the event. One of the main aims of the proposed framework would be to consolidate all the references of a particular event and map it to its corresponding EIIS maintained by the framework. However, due to the noisy nature of social media references an additional problem arise. We call it the problem of noisy reference.

- **Noisy Reference:** The three tweets in the example above are different in terms of information content related to the event. While the first one is very informative, the other two are surely related to the event, but not informative. The non-informative tweets might be interesting to the users having a personal conversation related to the event, but it has no value in conveying useful information about the event, that can act as an identity of the event. We consider these event related non-informative references as noise and do not want to map these tweets to the EIIS of the corresponding event.

Also, there might be a mismatch between the way an informative event reference is posted by an user and the way the proposed framework interprets the same in an automated fashion. This would result in loss of data integrity and increase the chances of erroneous results. Such problems are prevalent in the vanilla flavor of the EIIM as discussed in the previous section, giving rise to *false negative* and *false positive* errors.

Since the Event Identity Information Management framework operates in an open environment of social media as discussed earlier, two new scenarios leading to erroneous conditions are observed:

- **Noisy Event** - The first situation occurs when a reference gets mapped to an existing EIIS in the framework, although it refers to an event, which is out of the pre-specified list of events that are being tracked. Such a scenario is shown in Figure 2.5, where the event $E_p \in \Xi$ generates the reference $m_{x_{E_i}} \in M_{E_i}$, yet the external reference gets mapped to s_{E_1}.

  Spamming activity in social media channels can give rise to noisy events. For example, if we consider the following tweet:

  - *RT @BFDealz: http://t.co/TSJAigrVJI WHEELS SUPER TREASURE HUNT SUPERIZED HARLEY DAVIDSON FAT BOY LONG CARD 2014 #cpac2014 #sxsw*

  The above tweet might have been posted in the context of the event SXSW 2014 in order to market a Harley Davidson bike, but it also appears in the set of references for an unrelated event CPAC 2014. This is probably due to the fact that both the events took place parallely and the user posting this reference wanted to make it more visible by using the hashtags for both the trending events. This is often the case of spamming activities as discussed in Chapter 4. The tweet might not be related to both the events, yet it can get tracked by the framework.

- **Untracked reference** - The second situation occurs when there is a reference, which the framework is unable to track and map it to a pre-specified list of events although it refers to it. Such a scenario is shown in Figure 2.5, where the framework should have tracked the reference $m_{y_{E_3}} \in M_{E_3}$ and associate it with event $E_3 \in \Xi$, yet it is unable to do so and lose track of the reference in the process. This can happen due to the sampling bias as discussed in Chapter 4, which is one of the main challenges in collecting content from social media. The other reason could be errors in data collection process.

Therefore, the main broader problem that this dissertation solves is stated below.

Problem: *Given a pre-specified finite set of real-life events,* $\Xi = \{E_1, E_2, ..., E_n\}$ *generating a finite ordered stream of references* $M = \{m_{1_{E_1}}, m_{2_{E_1}}, ..., m_{i_{E_i}}, ..., m_{n_{E_n}}\}$ *in social media, and a finite set of EIIS structures* $S = \{s_{E_1}, s_{E_2}, ..., s_{E_n}\}$ *corresponding to each real-life event* $(E_i \in \Xi)$*, the problem is to resolve references of an event* (E_i)*, and to persistently extract, store and manage identity information of the event in its corresponding EIIS* s_{E_i}.

Towards the above objective, the dissertation seeks answers to the following questions:

- How to create the set of EIIS (S) for the pre-specified set of events of interest ?
- How to collect reference to the events containing identity information of the event from different social media platforms ?
- How to extract event identity information from the social media references ?
- How to integrate the event identity information from the unstructured references into the appropriate EIIS ?
- How to determine that new references from social media relates to an EIIS ?
- How to continuously integrate new information into the existing EIIS structures ?

Chapter 5 presents the design and implementation of the entire Event Identity Information Management framework that explains the strategies and novel techniques contributed by this dissertation in order to find answers to the above questions. A detailed survey of the existing literature related to the problem is presented in the next Chapter.

Chapter 3

Literature Review

3.1 Identifying High Quality Informative Content in Social Media

Identifying high quality content from the social media feeds that are related to events, is one of the main objectives of our research. As already discussed in Chapter 2, presence of spams, phishing, farm links, promotion of irrelevant content and development of nepotistic relationships are some of the major concerns of information quality in social media. Several effective solutions has been proposed in combating them by [31–34]. Among the different facets of information quality, credibility and trustworthiness of the references are also important. Due to the popularity and its ability to broadcast information at a tremendous pace, social media is also sometimes used by malicious users to spread misinformation and rumors [35]. In such cases, it becomes necessary to assess the credibility and trustworthiness of the information posted. It was showed by Castillo et al. [36] that selection of different types of features and automated classification based on supervised training can be used for detecting credible information about newsworthy topics in Twitter. In one of their works [37] they also proposed a general classification framework for identifying high quality social media content. They took into account the rich meta data like links between items and explicit quality ratings available in Yahoo! Answers website to train a supervised classification model. Credibility of events in Twitter was studied by Gupta et al. [38]. They used PageRank for propagating credibility scores on a heterogeneous network of events, tweets and users. They further constructed a graph between similar events and propagated the scores of the events from the previous network to estimate the credibility of other events. Ranking of tweets based on their credibility during trending events was proposed by Gupta and Kumaraguru [39]. They showed automated extraction of credible information from Twitter, by adopting supervised learning combined with relevance feedback approach using different features mined from tweets and the users posting them. Truthy[1], was developed by Ratkiewicz et al. to study information diffusion on Twitter and compute a trustworthiness score for a public stream of micro-blogging updates related to an event to detect political smears, astroturfing, misinformation, and other forms of social pollution [40].

Several mechanisms for ranking social media content in terms of their informativeness have been proposed. Ranking of microblogs like tweets are of particular interest to us as we consider tweets as a representative of short textual content produced in social media. 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[2]. 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[3]. A study of different state-of-the-art features and approaches commonly used for ranking tweets has been documented by [41, 42]. Seen[4] 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* [43] has been proposed by the scientific community for ranking tweets and users in Twitter [44–46]. TweetRank [46] is one such adaptation that ranks tweets by taking into account the direct relationships between tweets in the form of retweets and replies, as well as indirect follower-friend relationships, and usage of similar hashtags. Various learning to rank approaches have been used for ordering tweets retrieved for a given query in terms of their relevance and quality [47–49]. 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 [50]. 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.

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

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

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

[4] http://seen.co

Recently researchers have shown interest in investigating microblog summarization. Experiments have been conducted using both feature-based and graph-based approaches. However, in the context of our work only graph-based approaches are relevant. A comparison of different Twitter summarization algorithms was performed by [51]. Summarization of tweets for sporting events was performed by [52] using the phrase graph algorithm [53]. The popularly used graph-based summarization algorithms are *LexRank* [54] and *TextRank* [55]. Both the algorithms make use of the PageRank scheme of ranking homogeneous nodes in a graph constructed from the text that needs to be summarized and identify the salient text units for producing the summary. Our algorithm uses a similar technique for heterogeneous nodes. Our proposed framework also defines the semantics of the relationships between the nodes differently in the context of tweets. We use both *LexRank* and *TextRank* as evaluation baselines.

We propose implicit mutually reinforcing relationships between tweets, hashtags, text units, users and URLs forming a heterogeneous graph structure (*TwitterEventInfo-Graph*), which is novel and makes our work different from any prior work (refer Chapter 5). Scores are assigned to the association between the nodes representing the semantics of their relationships. We implement an iterative algorithm (*TwitterEventInfoRank*) for ranking the nodes of the graph and propagating the event-specific scores of the nodes to its neighboring nodes based on the measure of their association. To our knowledge, this is the first work that identifies novel relationships between different units of content in Twitter and implements a graph-based algorithm for ranking them simultaneously in the context of an event.

3.2 Entity Resolution

Entity resolution has been known for more than five decades as the record linkage or the record matching problem in the statistics community [56–58]. In the database community, the problem is defined as merge-purge [59], data de-duplication [60, 61], and instance identification [62]. In the Artificial intelligence community, this problem is described as database hardening [63], and name matching [64]. The names co-reference resolution, identity uncertainty, and duplicate detection are also commonly used to refer to the same task [65]. The term Entity Resolution (ER) first appeared in publications by researchers at the Stanford InfoLab led by Hector Garcia-Molina and is defined as the process of identifying and merging records judged to represent the same real-world entity [66]. In the context of the work presented in this thesis a pre-defined real-life event is considered as an entity. For detailed definition of an event please refer Chapter 2.

Despite the differences in nomenclature used by these authors, the ER process actually comprises five major sub-tasks or activities [29] which are

1. *Entity reference extraction* – locating entity references in unstructured textual information.
2. *Entity reference preparation* – profiling, standardizing, cleaning, and enhancing reference information in preparation for resolution.
3. *Entity reference resolution* – the process or algorithm for determining when references are equivalent, often through direct matching of attributes.
4. *Entity identity management* - creating and maintaining persistent data structures that represent the identities of external entities, the focus of the proposed research.
5. *Entity relationship analysis* – exploring relationships among distinct entities such as household relationships or shared communication.

The *Event Identity Information Management* Life Cycle (Chapter 5) as proposed in this thesis reflects and implements all of the above activities. Historically the focus of ER research has been on Activity 3, the methods for carrying out the resolution process itself. The majority of published research literature falls into this area. The first formal model for resolution was the Fellegi-Sunter Model of Record Linkage [56], which uses a decision-theoretic approach establishing the validity of principles first used in practice by Newcombe [57]. This was followed by the Stanford Entity Resolution Framework (SERF) developed at the Stanford InfoLab [67]. The SERF Model formalizes the generic ER problem as the interaction of two functions for comparing and merging records as black-boxes and defines the conditions required for these functions to give a unique ER result. It also formulates a family of so called "Swoosh" algorithms (G-Swoosh, R-Swoosh, and F-Swoosh) for carrying out the ER process. With the rise of big data a distributed algorithm D-Swoosh [68], was also proposed that can be implemented in a big data environment. More recently the Talburt-Wang Algebraic Model of ER has been proposed [69] that views ER as a problem of partitioning a given set of references.

In addition to research on Activity 3, there has also been extensive research in the area of information extraction (IE) that is directly related to the ER Activity 1, reference extraction. The task of entity extraction is also more relevant to social media, due to the unstructured nature of the content. One of the main emphases in the realm of unstructured textual content for last two decades has been in the task of extracting named entities and categorizing them into types. Competitions like MUC (Message Understanding Conference), CoNLL (Conference on Computational Natural Language

Learning) and ACE (Automatic Content Extraction) spearheaded the development of new techniques in this domain. This led to the development of sophisticated tools like Stanford NER [70], OpenNLP [71], GATE [72], LingPipe [73] and NLTK [74]. Variety of techniques ranging from hand-coded rules, automatic rules, to statistical machine learning techniques like hidden Markov models, maximum entropy and conditional random fields have been proposed. A comprehensive survey of the techniques could be found in [75, 76]. A study of various efforts in extracting information from micro-blogs could be found in [77] and a survey of named entity recognition and classification could be found in [78]. Efforts have been made by the industry in building crowd sourced knowledge bases like freebase [79] and dbpedia [80] for the purpose of entity extraction. A recent effort from the industry for extracting entities from social media and building scalable knowledge bases for doing so has been documented in [81, 82]. The rise of online social networks, has also motivated new research into the ER Activity 5, entity relationship analysis [83]. With the rise of big data, the modern trend is to perform entity resolution process in humongous volumes of data and scale it horizontally [84, 85]. In spite of the recent efforts in the field of entity extraction and resolution from unstructured text, there is no generic framework that solves the problem of persistently collecting and managing entity identity information from social media. The development of Event Identity Information Management from social media is a pioneering effort in the field of entity resolution and would create new avenues of research.

Traditionally, entity identity resolution and management (Activity 4) has been a subject of system administration and management of user identities in large organizations. For the first time [1], showed the intersection of identity management, master data management and entity resolution could be used for managing identities of real-life entities in information systems, that could further play an important role in data integration and information quality. Entity identity management in social media mainly comprises of resolving and integrating profiles of the same person in social networking websites. The FOAF project has been playing an important role in all such efforts [86–88]. A very nice endeavor has been made by the OKKAM project for integrating and managing the multiple entity identifiers in various knowledge bases across the Internet [89]. To our knowledge, we are the first to propose a framework for collecting and extracting identity information of events from social media and use the concepts of entity identity management and entity resolution for persistently managing their identities with respect to time.

3.3 Event Identification in News Text

The event detection task [90] 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 [15, 16]. While others explored the detection of past events from archived news documents [91].

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. [92] 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 [93] 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. [94] 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. [95] 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.

Extracting events from text has been the focus of numerous studies as part of the NIST initiative for Automatic Content Extraction (ACE) [96, 97]. 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 [96] divided the event extraction task into different subtasks, including identification of event keyword triggers, and determination of event coreference, and then used machine learning methods to optimize and evaluate the results of each subtask. Ji and Grishman [97] 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. [98] 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 always 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 [99]. For example, named entity recognition methods typically show 85-90% accuracy on longer texts, but 30-50% on tweets [100]. Therefore, new approaches had to be taken, leading to new techniques for detecting events in social media, which we discuss next.

3.4 Event Identification in Social Media

Identification of events and event related content from social media is still in its infancy and needs to be studied more. Several related papers explored the unknown event identification scenario in social media. Weng and Lee [101] proposed wavelet-based signal detection techniques for identifying real-life events from Twitter. These techniques can detect significant bursts or trends in a Twitter data stream. Sankaranarayanan et al. [102] identified late breaking news events on Twitter using clustering, along with a text-based classifier and a set of handpicked news seeders. But they do not take into account the filtering of non-event content, which results in poor performance. Segregating the messages that have high likelihood of containing event related informative content from the ones with chances of having non-informative content, or content that are not at all related to an event are at the core of the work presented in this thesis. Petrovic et al. [103] used locality-sensitive hashing to detect the first tweet associated with an event in a stream of Twitter messages. Rattenbury et al. [104] analyzed the temporal usage distribution of tags to identify tags that correspond to events. Chen and Roy [105] 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. Becker et al. [106] defined multi-feature similarity metrics based on the textual and non-textual features associated with the social media documents in order to automatically identify events and their related content. They use the general text-based classifier suggested in [102] and a method for identifying top events suggested by [103] as baseline approaches in their evaluations and achieved better precision scores.

New techniques have been proposed recently for identification of known 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 [107, 108]. Sakaki et al. [107] 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. Benson et al. [109] 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. Most of these approaches are tailored towards one specific social media site. Becker et al. [110] extracts event features, that are often noisy and missing and use them to develop query formulation strategies for retrieving content associated with a planned event from Twitter [111] as well as different social media websites [110].

Our method of tracking events is similar to the idea of identification of known events. We also use predefined hashtags and query words to bootstrap the process of collecting data related to a known set of events. However, we introduce and implement the concept of Event Identity Information Structures that are mapped in a one-to-one mapping with the events that we track. The Event Identity Information Structures persistently stores information that acts as identity of an event as the event evolves with time. This identity information is further processed and ranked in order to identify the top event-specific informative units that is further used for tracking new event related content being generated in different social media channels. Also the emphasis of our research is more on information quality, which is absent in most of the previous research in social media. Instead of just identifying event related content, we identify event-specific informative content. Also, the technique that we develop for identifying event-specific informative content from microblogs (Twitter) leverages hashtags, text units, users, posts and URLs. All these metadata are available in most of the social media websites producing short textual content. Therefore our technique should be applicable to other such platforms. We plan to explore it in the future.

Chapter 4

Challenges in Mining Event Related Content from Social Media

The characteristics of social media websites as discussed above in Section 2.2 makes the domain extremely unstructured and uncontrolled. This gives rise to different challenges in mining information from all types of social media platforms. Some of the major challenges applicable to the context of this dissertation are discussed below. The solutions to these challenges as proposed in this dissertation is also refered while explaining them. Social media posts from the datasets collected for the experiments in Chapter 5 (Section 5.2) are presented as examples while discussing the problems.

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 [112]. They suffer from the dilemma of assessing the accuracy and quality of information content in the sources being produced over their freshness. Thus,

[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/

collecting different types of references of events from social media, assessing their quality, resolving and extracting identity information of the events poses great challenges in such a situation.

For example, 284 million monthly users of Twitter posting 500 million tweets per day produces a variety of content[4]. 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. In this dissertation, we call them as event-specific informative tweets. Table 4.1 presents some examples of different types of tweets shared during real-life events. One of the main problems due to information overload is to identify these tweets among millions of tweets being produced during the event.

We develop techniques in this dissertation that helps in overcoming this challenge. In Chapter 5 (Section 5.4), we develop a generic supervised classifier for Twitter in order to identify event references in real-time from Twitter that has high likelihood of containing high quality information. This results in segregation of informative references from the non-informative ones, and filtering the informative references for further processing. The *EventIdentityInfoRank* algorithm implemented in Chapter 5 (Section 5.7.2), processes the content in the filtered tweets and results in a ranked list of tweets that has high quality event-specific informative content.

4.2 Veracity of Sources

Judging the accuracy of the information and detecting relevant, event-specific informative content from social media constitutes another challenging situation due to the malevolent practices of spam users. For trending topics the search engines have started

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

Table 4.1: 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'***

showing real-time feeds from social media websites in their search results. This has attracted spammers who post trending hashtags or keywords along with their spam content in order to attract people to their websites offering products or services [31]. For example, the tweet (*RT @BFDealz: http://t.co/TSJAigrVJI WHEELS SUPER TREASURE HUNT SUPERIZED HARLEY DAVIDSON FAT BOY LONG CARD 2014 #cpac2014 #sxsw*) was posted during the parallel occurrence of CPAC 2014 (a political conference) and SXSW 2014, but has nothing to do with the events. Instead it leads to a deal related to Harley Davidson bike promoted using popular event related hashtags #cpac2014 and #sxsw.

An alarming 355% growth of social spam has been reported in 2013[5]. Social media has also been instrumental in spreading misinformation and rumors. Spread of misinformation not only results in pandemonium among the users[6] but also result in extraction of completely wrong information about events. The users in the social media websites also develop nepotistic relationships in order to get higher scores in the ranking techniques with malicious intentions [113]. This also helps them to spread spam and other malicious content. Such behavior can also lead to cyber attacks.

Some examples of spam tweets in our collected dataset are shown in Table 4.1. Most of the existing techniques face problems in combating spam. *EventIdentityInfoGraph* (Section 5.7.1) is presented and explained in this dissertation that defines mutually reinforcing chains in Twitter for identifying the most event-specific informative references and filters out the spam tweets after ranking its nodes using *EventIdentityInfoRank* (Section 5.7.2). The algorithm can also identify users who are producing event-specific

[5]http://www.likeable.com/blog/2013/11/10-surprising-social-media-statistics/
[6]http://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter

informative content and URLs (images, videos, news articles) that are extremely relevant to the event.

4.3 Multiple Data Sources with Variety of Content

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.

Although, the type of data and the format of the data returned by the social media services are different, yet most of them have certain common meta-data associated with messages obtained in the responses to the API requests. These are hashtags, text units extracted from the messages/descriptions, the messages itself, the users posting them and the URLs that lead to external sources of related information. Some of the websites where hashtags are not present, the associated tags can be used instead. Table 4.2 shows the presence of these meta-data in the most popular social media platforms that carry event related information.

Table 4.2: Presence of different meta-data in popular social media websites.

Social Media Website	**Hashtags**	**Users**	**Text Units**	**Messages/ Description**	**URLs**
Facebook	Yes	Yes	Yes	Yes	Yes
LinkedIn	Yes	Yes	Yes	Yes	Yes
Pinterest	Yes	Yes	Yes	Yes	Yes
Instagram	Yes	Yes	Yes	Yes	Yes
Twitter	Yes	Yes	Yes	Yes	Yes
Flickr	Yes	Yes	Yes	Yes	Yes
Google Plus	Yes	Yes	Yes	Yes	Yes
YouTube	Yes	Yes	Yes	Yes	Yes

The developed techniques that are explained in this dissertation rely only upon the above meta-data. This makes the techniques generic and useful for most of the social media channels.

4.4 Informal Text

Unlike sources of news media and edited documents on the web, the textual content of the social media references 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 [114]. 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 [115]. 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 [99]. For example, named entity recognition methods typically show 85-90% accuracy on longer texts, but 30-50% on tweets [100]. 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.

In order to solve some of these problems we make use of additional resources in this dissertation that are compiled by us. Some of these resources are list of slang words, list of acronyms and list of stop words commonly used in short social media text (refer Chapter 5). We use these resources in order to aid the natural language processing techniques so that it becomes capable of extracting the required information with higher accuracy. We also experimented with a state-of-the-art entity extraction service AlchemyAPI[7] (Chapter 6) and found that the final results obtained using our data processing pipeline are better.

4.5 Searching for Information in Long Tail

Due to the power law distribution of the Internet [116], and the present search engine technology, the 'Short Head' is generally dominated by the mainstream media websites. As illustrated in Figure 4.1 the top 10 search results for "Egyptian Revolution",

[7]http://alchemyapi.com

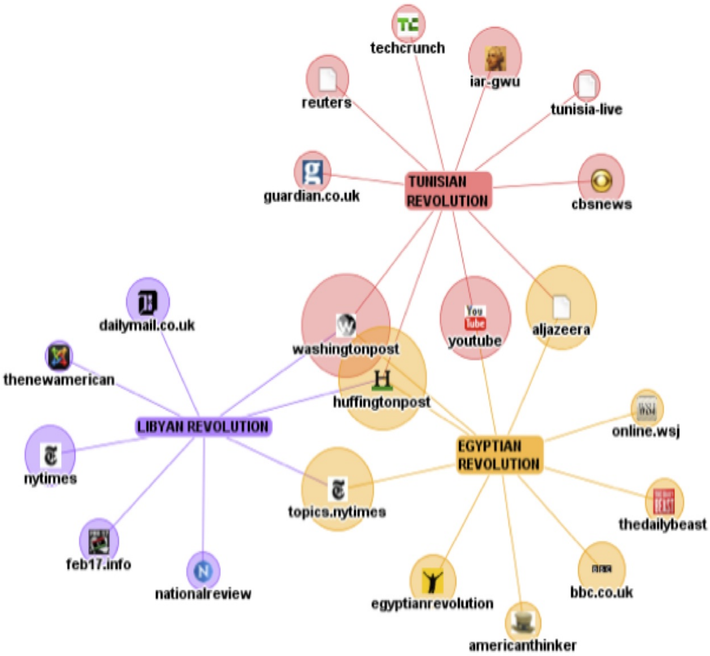


FIGURE 4.1: Top 10 Google search results for "Egyptian Revolution", "Libyan Revolution", and "Tunisian Revolution", visualized using TouchGraph.

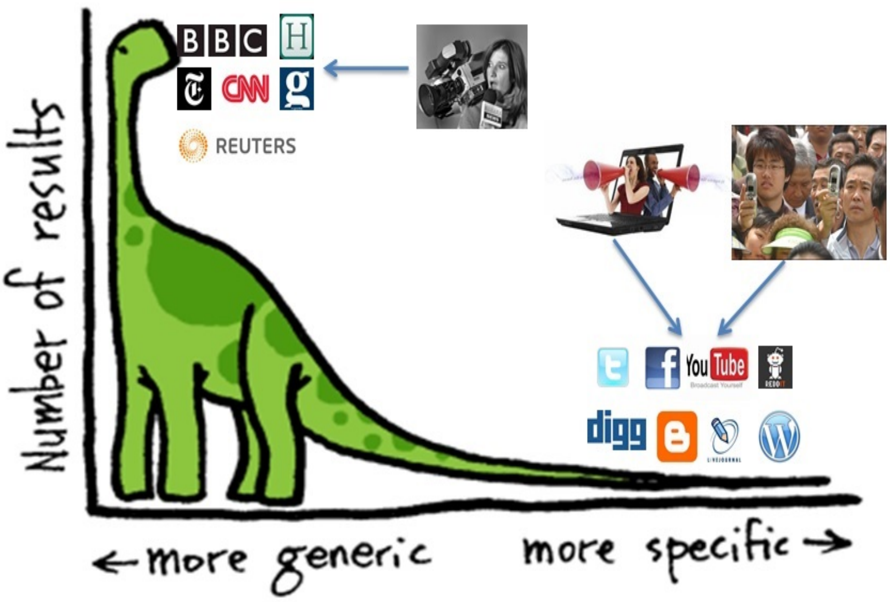


FIGURE 4.2: Short Head Vs Long Tail media sources.

"Libyan Revolution", and "Tunisian Revolution" by Google, visualized using Touchgraph[8], mostly retrieved mainstream media references. Consequently, the social media sites get buried in the Long Tail [117] as shown in Figure 4.2. However, references from the social media channels, act as hubs of specific information about real-life events [118]. On the other hand, the mainstream media sources often gloss over the intricate details

[8]http://touchgraph.com

while covering a real-life event. They are often biased, regulated by the government, and may not portray the true picture of an event [19]. While, social media references often contain unbiased, uninhibited, and unedited opinions from people. Blogs have been accepted as more credible sources of information over mainstream media references by the weblog users [119]. Thus the references, 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 [6, 14, 120], and more.

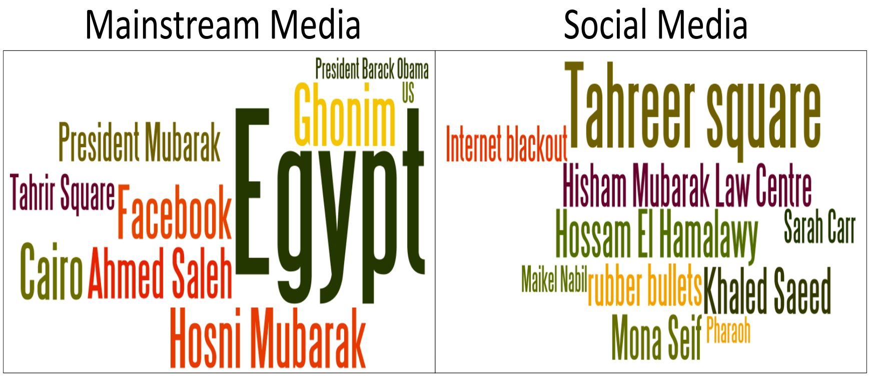


FIGURE 4.3: Top 10 entities from mainstream media and blogs.

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 Figure 4.3. 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.

A person interested to know about an event in detail, may miss out the novel and specific information available in social media by relying on the top results from the popular search engines. It is a challenge for the current ranking schemes to retrieve the event-specific social media content from the long tail. Moreover, in the words of Chris Anderson [121], *"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 also motivated us to look for techniques in this dissertation, that would help in identifying these otherwise buried sources providing highly specific information related to an event.

4.6 Sparse Link Structure

The casual nature of the users posting content in social media channels gives rise to the challenge of sparsity in link structures. Most often the users who posts content, do not provide links to the original source of information. Also, the behavior of linking to other similar content or building citation networks between information posted about the same topic is completely absent among the social media users. This creates an extremely sparse link structure between the user-generated posts. This further creates problems for the traditional and state-of-the-art searching techniques such as PageRank [122] that performs well in ranking web pages.

In this dissertation, we define novel implicit relationships, intrinsic to content available in the social media channels for solving this issue, and rank them showing better performances than some of the popular baseline techniques.

4.7 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[9]. 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.

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

4.8 Lack of Evaluation Datasets

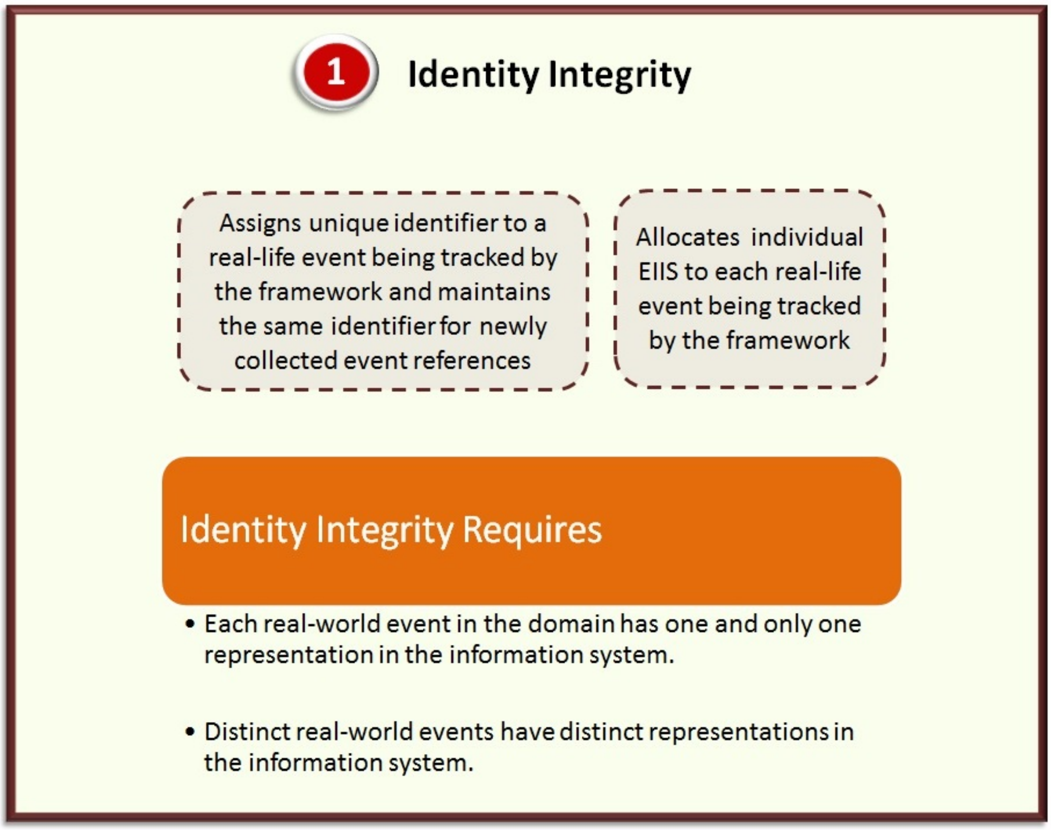
There is a lack of ground truth evaluation data for most of the social media text mining tasks. In traditional data mining research, there is often two types of datasets. One of them is known as training dataset and the other is known as test dataset. The models are trained or developed using the training 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.

In this dissertation we find novel ways to validate our results. We rely on both automated as well as manual method of evaluations, as considered appropriate in different scenarios. We use independent annotators for all the annotation tasks, educate them about the tasks and the required event related information. We also report the percentage of agreement in their performed tasks, that lies above the accepted thresholds (Chapter 6).

Chapter 5

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

FIGURE 5.1: Identity Integrity component of the EIIM life cycle.

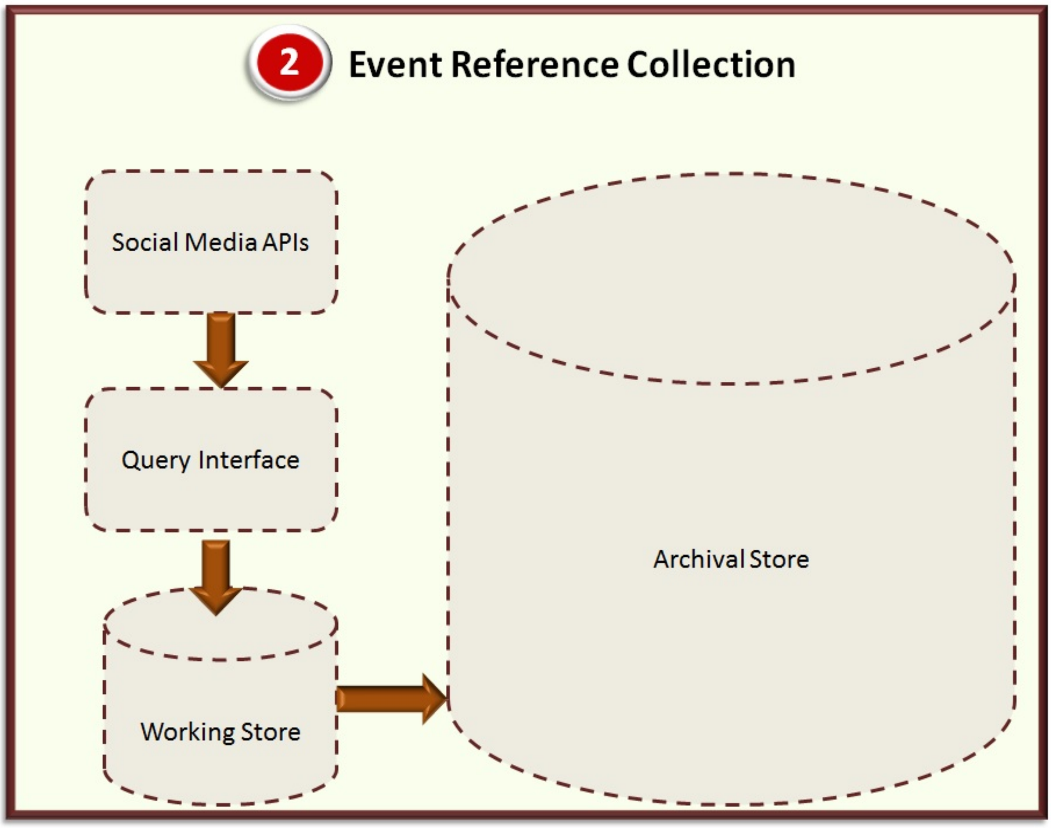


5.1 Identity Integrity

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.

5.2 Event Reference Collection

FIGURE 5.2: 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. A query interface is implemented that allows an user of the system to pass query parameters (for example event related hashtags and key words) that bootstraps the data collection and event tracking process. As already shown in Table 4.2 most of the popular social media channels allow hashtags, the data for the experiments were collected using a popular hashtag for the respective events.

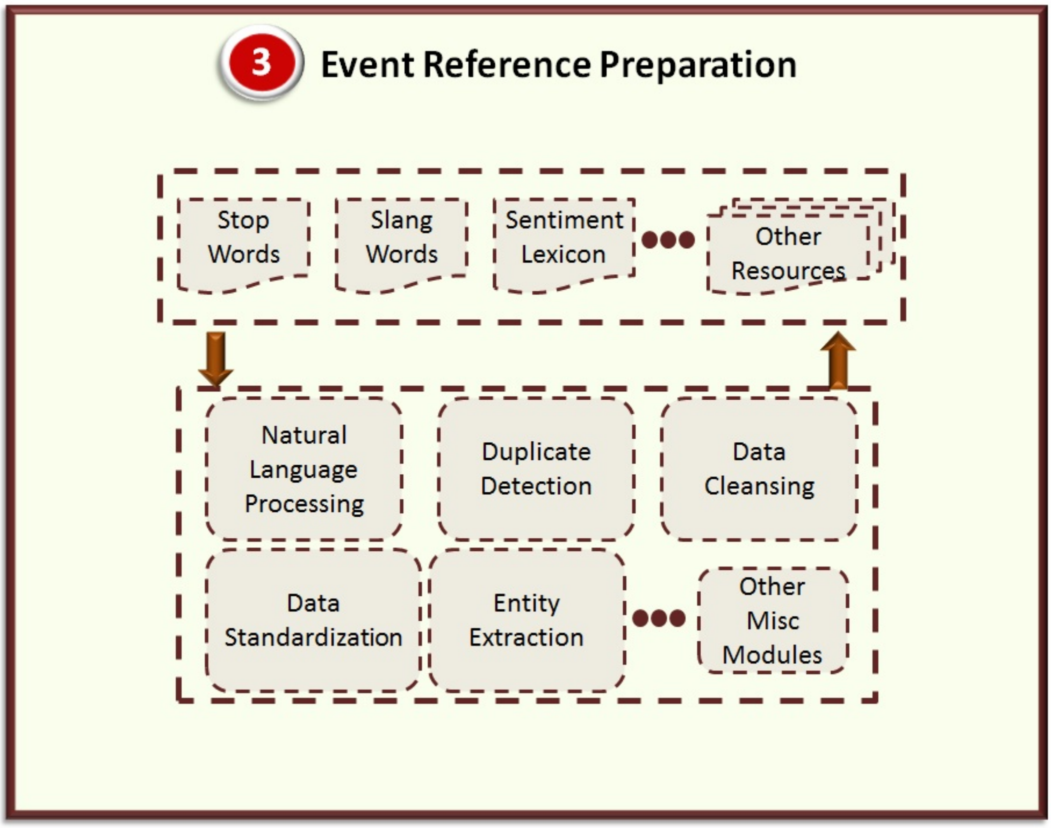
Due to extreme popularity of Twitter, data from it was collected for representing the microblog genre and short textual social media references. Four million tweets (approx) related to five different events were collected. Details of the collected event references are provided in Table 5.1. The tweets were collected over the given period of time, by providing a popular hashtag to the Twitter streaming API as shown in Table 5.1 (for details about Twitter Data Collection please refer Appendix A). Only English language tweets are considered for the experiments as the available natural language toolkits performs well for English, and the annotators used for different tasks are only proficient in English language.

TABLE 5.1: 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

5.3 Event Reference Preparation

FIGURE 5.3: Event Reference Preparation component of the EIIM life cycle.



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 references in order to make them suitable for further processing by the other components of the EIIM life cycle. Several resources are compiled in order to tackle the challenges posed by short informal text prevalent in social media.

The tweets collected using the previous component goes through the following preprocessing steps:

5.3.1 Parts-of-speech tagging

The Natural Language Toolkit (http://nltk.org) POS tagger is used for tagging the raw text of the tweets. All the words of a tweet is assigned one of the following parts-of-speech:

- Noun

- Adjective
- Verb
- Adverb
- Preposition
- Interjection
- Pronoun
- Article

The tagged tweet is also separately stored and maintained. These tags are used later in different components down the pipeline. The Penn Treebank tags[1] are used for tagging and is parsed accordingly for identifying words with a certain parts-of-speech.

5.3.2 Special Character Detection

All the special characters that are not alphanumeric are detected and the total number of special characters in a tweet is stored.

5.3.3 Data Cleansing

The raw text of the tweet is extracted and cleaned. All the user mentions, hashtags, retweet symbols, URLs and special characters are removed and the entire tweet is converted into lowercase.

5.3.4 Duplicate Detection

The tweets after cleaning are assigned a md5 hash code, which helps in detecting duplicate content. Tweets having the same hash code are considered to be redundant copies of each other, and only a single copy of the tweet is finally stored in the database. This technique also helps in detecting the retweets of a tweet that contain the same content. Also there are certain tweets that are shared with the same content but has different user mentions and with different combination of the words expressing the content. For example, the following tweets talks about the same video shared by mashable and does not present any new information.

[1] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

- RT @mashable: Timelapse video reveals massive size of New York City protests http://t.co/zhqHpkDLk1 #MillionsMarchNYC http://t.co/WktxssAfDp
- RT @dianebhartford: "@mashable: Timelapse video reveals massive size of New York City protests http://t.co/CE0VIyHnLe #MillionsMarchNYC htt...

After the data cleansing step, the duplicate detection scheme identifies both the tweets to be same and only maintain a single copy. This process occurs in real-time. Whenever, a new tweet is obtained from the straming API, the md5 hashcode is calculated after going through the previous data pre-processing steps. A hashtable is maintained in the memory that is constantly searched for the presence of the generated hashcode. If the hashcode is already present then the tweet is dropped and not stored.

5.3.5 Stop Word Detection and Elimination

A list of English stop words is compiled that is publicly shared in the following URL :

- https://github.com/dxmahata/EIIMFramework/blob/master/CodeBase/EventIdentityInformationManagement/Resources/englishStopwords.txt

This list is used for detecting the stop words in English language tweets. The stop words are eliminated and the number of stop words detected is recorded.

5.3.6 Slang Word Detection

Slang words commonly used on the Internet and twitter specific slang publicly shared by FBI[2] is combined together for compiling a list of English slang words. This list is used for detecting and extracting the slang words from the tweets. The number of slang words detected is recorded. This list is also used later to detect the slang hashtags and slang text units.

The compiled list of twitter specific slang words is publicly shared and can be obtained from the following URL :

- https://github.com/dxmahata/EIIMFramework/blob/master/CodeBase/EventIdentityInformationManagement/Resources/slangWords.txt

[2]https://www.documentcloud.org/documents/1199460-responsive-documents.html#document/p1

5.3.7 Feeling Word Detection

A list of words expressing feelings on the Internet, obtained from wefeelfine.org is used for detecting and extracting the feeling words from a tweet. The number of feeling words detected is recorded and the extracted feeling words are stored. This list is also publicly available and can be obtained from the following URL :

- https://github.com/dxmahata/EIIMFramework/blob/master/CodeBase/EventIdentityInformationManagement/Resources/feelingWords.txt

5.3.8 Tokenization

The tweet obtained after performing the data cleansing steps and elimination of stop words are tokenized into unigram and bigram tokens using the tokenizer module available in NLTK. The set of tokens thus obtained are stored separately.

5.3.9 Stemming

The unigram tokens obtained after tokenization are stemmed and a separate list of stemmed tokens are stored. A standard Porter stemmer available with NLTK library is used for the purpose.

5.3.10 Tweet Meta-data Extraction

Several meta-data that are associated with each tweet obtained from the JSON response of the streaming API are extracted. Some of these meta-data are :

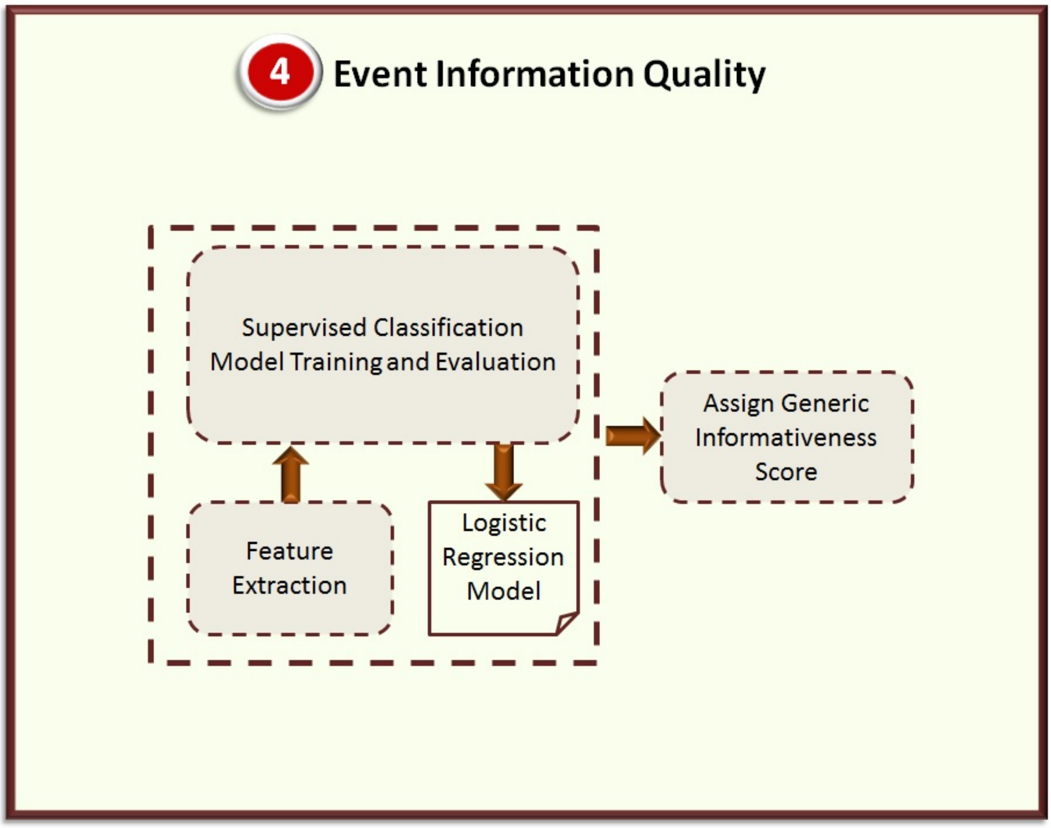
- Expanded URLs
- Hashtags
- Retweet Counts
- Favorite Counts
- User Mentions
- User Follower Counts
- Verification Information
- Time Information

5.3.11 Named Entity Extraction

Named entities such as name of persons, animals, places, cars and organizations are extracted from the raw tweets. For this purpose the entity extraction service of AlchemyAPI[3] is used.

5.4 Event Information Quality

FIGURE 5.4: Event Information Quality component of the EIIM life cycle.



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 quality information. In order to make a generic module for identifying high quality event related informative references we implemented a logistic regression classifier. Once the classifier is trained, it is used for assigning generic informativeness score to the tweets in real-time as they are collected using the streaming API. The component aids in solving the problem of information overload. Just like an user searching for relevant informative

[3]http://alchemyapi.com

content about an event faces the challenging situation of information overload as discussed in Chapter 4, Section 4.1, it is also a challenge for automated systems to process the huge amount of content coming at high velocity and extract useful information out of it. By filtering out the tweets that are less likely to contain useful and high quality information it solves the problem of information overload for the other components of the EIIM framework.

5.4.1 Annotated Dataset

We use a publicly available annotated dataset from the CrisisLex[4] website shared by Olteanu et al. [123]. The collection includes tweets collected during 26 large crisis events in 2012 and 2013, with about 1,000 tweets labeled per crisis for informativeness (i.e. "informative", or "not informative"), information type, and source. 28,000 tweets (about 1,000 in each collection) were labeled by crowdsourced workers according to informativeness (informative or not informative), information types (e.g. caution and advice, infrastructure damage), and information sources (e.g. governments, NGOs).

For example, for the Colorado wildfire[5] event, the following tweets were assigned labels of "related and informative", "related but not informative", and "not related", respectively.

- *Related and Informative* - #Media Large wildfire in N. Colorado prompts evacuations: Crews are battling a fast-moving wildfire http://t.co/ju1BGTKH #Politics #News
- *Related but not Informative* - RT @LarimerSheriff: #HighParkFire update http://t.co/hBy5shen
- *Not Related* - #Intern #US #TATTOO #Wisconsin #Ohio #NC #PA #Florida #Colorado #Iowa #Nevada #Virginia #NV #mlb Travel Destinations ; http://t.co/TIHBJKF2

Not all the tweets were in English language. We selected English language tweets for training the logistic regression model. There were only 9729 tweets.

5.4.2 Feature Selection and Training

In order to train the model we assigned a score of 1 to the tweets that were labeled 'related and informative', and all the other tweets labeled as 'related-but not informative',

[4]http://crisislex.org/data-collections.html
[5]http://en.wikipedia.org/wiki/2012_Colorado_wildfires

and 'not related' were assigned a score of 0. The choice of features was governed by previous works related to identifying high quality information from Twitter [36, 39, 49, 124]. The list of features selected for the model are:

1. **Has URL** - has a value of 1 if the tweet contains URL or else has a value of 0.
2. **Number of Words** - total number of unigram tokens extracted from the raw tweet text.
3. **Number of Stop Words** - total number of English stop words detected in the raw tweet.
4. **Number of Feeling Words** - total number of feeling words detected in the raw tweet.
5. **Number of Slang Words** - total number of slang words detected in the raw tweet.
6. **Number of Hashtags** - total number of hashtags used in the raw tweet.
7. **Number of User Mentions** - total number of user mentions detected in the raw tweet.
8. **Tweet Length** - total number of characters used in the raw tweet.
9. **Unique Characters** - total number of unique characters used in the raw tweet.
10. **Special Characters** - total number of special characters detected in the tweet.
11. **Favorite Count** - total number of favorite count of the tweet at the time it was collected.
12. **Retweet Count** - total number of retweet count of the tweet at the time it was collected.
13. **Verified** - has a value of 1 if the user posting the tweet is a verified user by Twitter or else has a value of 0.
14. **Number of Nouns** - total number of nouns detected in the tweet, without considering the hashtags and the user mentions whenever they are tagged as nouns.
15. **Number of Adjectives** - total number of adjectives detected in the tweet, without considering the hashtags and the user mentions whenever they are tagged as adjectives.
16. **Number of Verbs** - total number of verbs detected in the tweet, without considering the hashtags and the user mentions whenever they are tagged as verbs.

17. **Number of Adverbs** - total number of adverbs detected in the tweet, without considering the hashtags and the user mentions whenever they are tagged as adverbs.

18. **Number of Pronouns** - total number of pronouns detected in the tweet, without considering the hashtags and the user mentions whenever they are tagged as pronouns.

19. **Number of Interjections** - total number of interjections detected in the tweet, without considering the hashtags and the user mentions whenever they are tagged as interjections.

20. **Number of Articles** - total number of articles detected in the tweet, without considering the hashtags and the user mentions whenever they are tagged as articles.

21. **Number of Prepositions** - total number of prepositions detected in the tweet, without considering the hashtags and the user mentions whenever they are tagged as prepositions.

22. **Formality** - which is defined as follows, Formality = (#nouns + #adjectives + #prepositions + #articles - #pronouns - #verbs - #adverbs - #interjections + 100)/2 and is proposed in [125]. #nouns, denotes the number of nouns detected in the tweet, and so on.

TABLE 5.2: 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%		

5.4.3 Model Evaluation

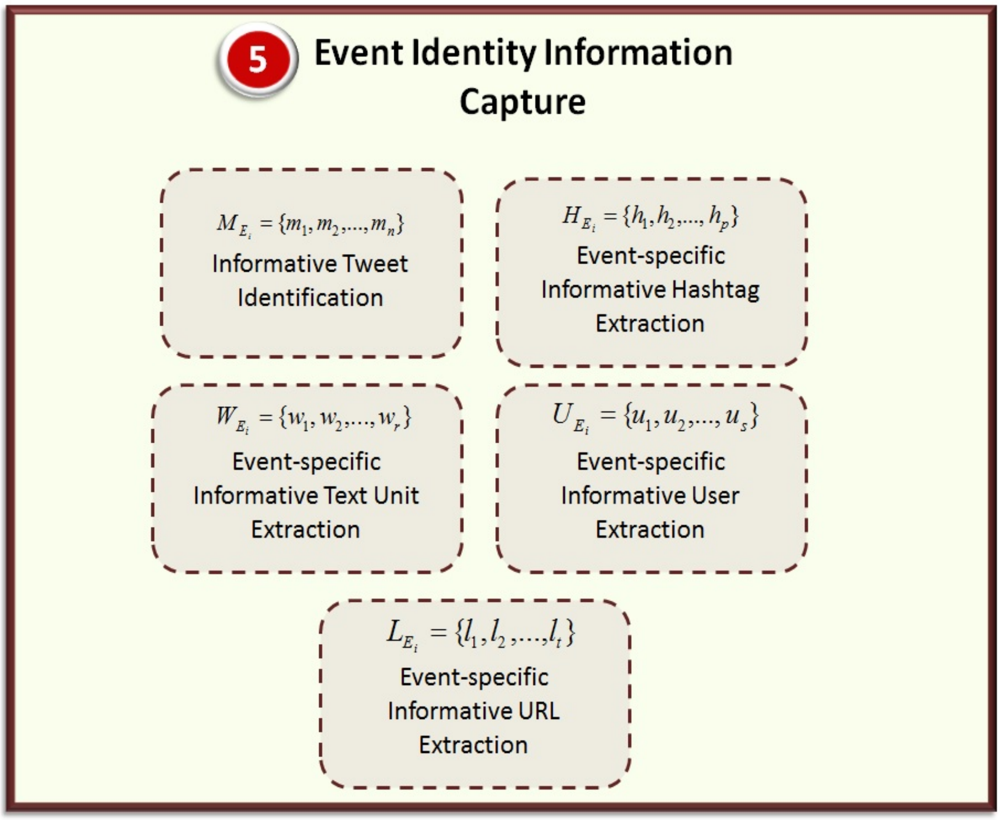
10-fold cross validation was performed, with 'l1' penalty, resulting in a model with an accuracy of 76.64%. Table 5.2 lists the evaluation measures obtained while training the classifier. The ROC AUC Score of the classifier is 0.6934.

5.4.4 Assignment of Generic Informativeness Score

The 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.

5.5 Event Identity Information Capture

FIGURE 5.5: Event Identity Information Capture component of the EIIM life cycle.



The main functions of this component are:

- This component 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 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 following three events.

- CPAC 2014.
- SXSW 2014.
- Sochi Winter Games 2014.

The analysis and the conclusions we made from it, is presented next.

5.5.1 Content Analysis of Event Related Tweets

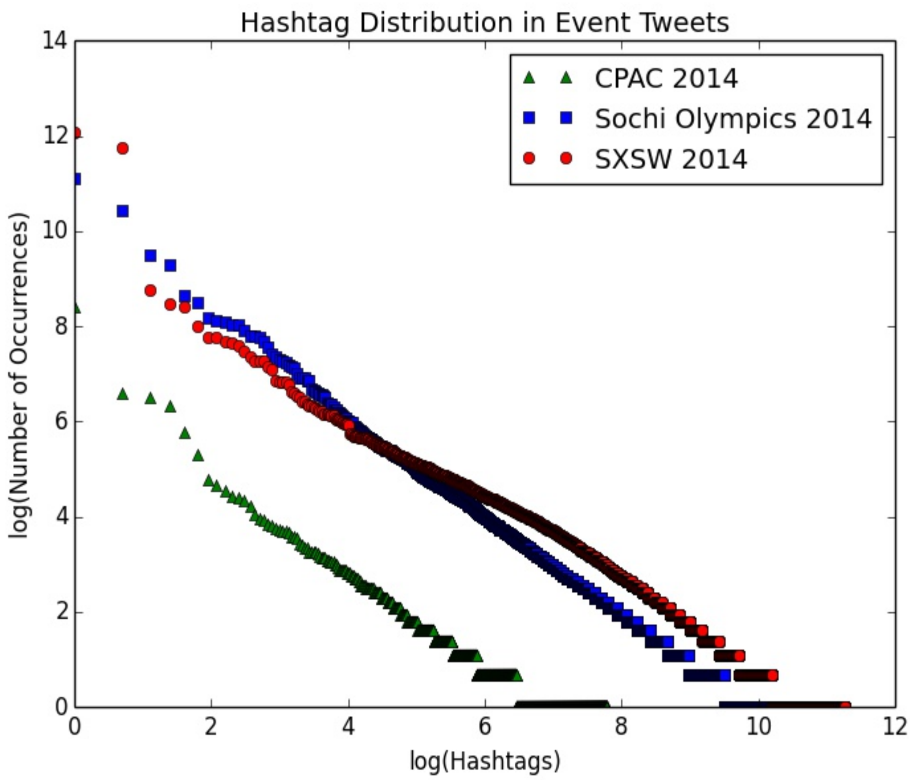
Details of the data collected for the analysis are provided in Table 5.1. The data collection task was accomplished by 'Event Reference Collection' component and was then preprocessed by the 'Event Reference Preparation' component.

Twitter allows its users to post short messages with a limitation of 140 characters. Users not only post plain textual content in their messages but also share URLs, linking to other external websites, images and videos. The images and videos are labeled as media elements by Twitter. Apart from curating new content, the users also share content produced by others. This activity is known as *retweeting*, and such tweets are preceded by the special characters *RT*. The messages are normally written by a single person and are read by many. The readers in the context of Twitter are known as *followers*, and the user whom the other users 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 conversation with them.

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. #icwsm2015). They are widely used by the users in order to add context to the tweets, 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 [126]. When tweeting about real-life events the users also tend to use hashtags in order to post event-specific content. For example '#Egypt' and '#Jan25', were among the most popular hashtags in Twitter used for spreading, organizing and analyzing information related to 'Egyptian Revolution of 2011' [127].

Given the mechanisms of user interactions and content production in Twitter, we started our analysis with the assumption that the content of a tweet is primarily composed of hashtags, words for expressing and conveying information, and URLs that lead to additional information about the content. We plotted the distribution of occurrences of all the hashtags, tokens and shared URLs for each event. Due to the short length of the tweets we only considered unigram tokens. We also plotted the distribution of the number of tweets posted by the users. We observed a power law distribution for all of them (refer Figures 5.6, 5.7, 5.8 and 5.9). This gave us an intuition that there are skewed sets of hashtags and tokens that are widely used for posting content related to an event. There is also a specific set of URLs that become popular in comparison to others and a set of users who are more active than others in posting event related content.

FIGURE 5.6: Distribution of hashtags in event related tweets.



Our second step was to use the logistic regression model developed for the 'Event Information Quality' component and assign informativeness 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. We calculated the average values of different content characteristics of the tweets. Top ten percent of the frequently occurring hashtags and nouns were considered as top hashtags and top nouns

FIGURE 5.7: Distribution of tokens in event related tweets.

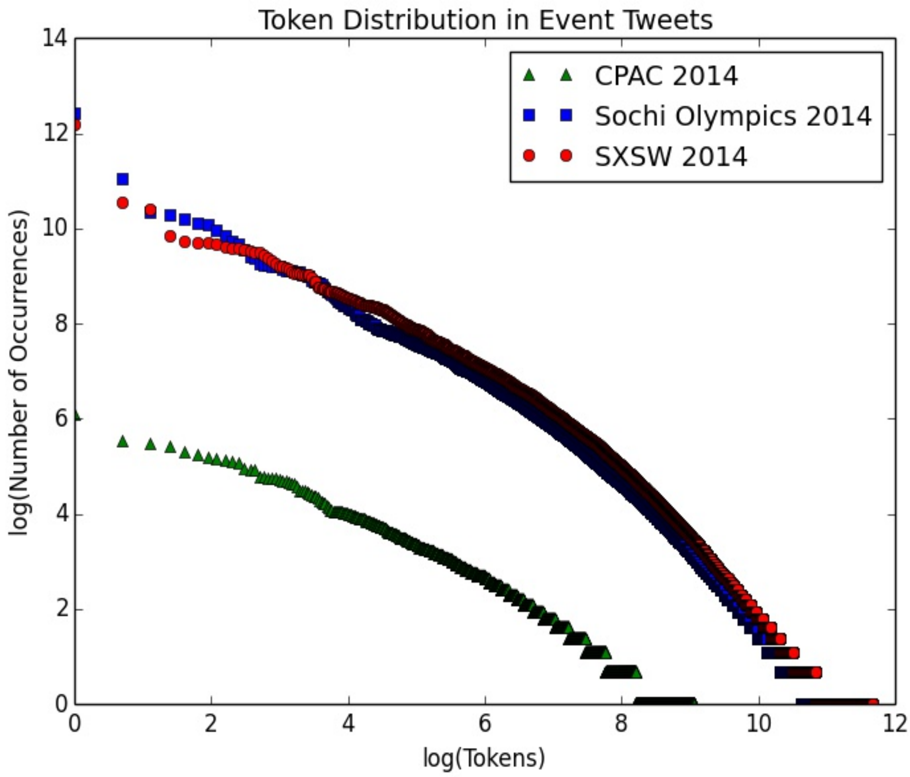


FIGURE 5.8: Distribution of URLs in event related tweets.

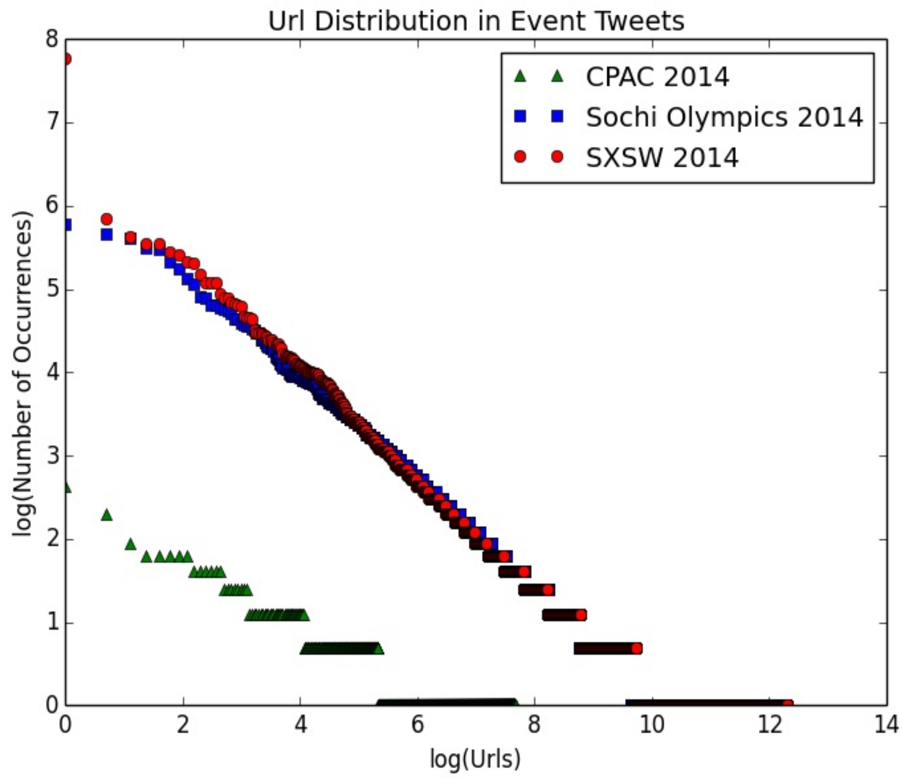
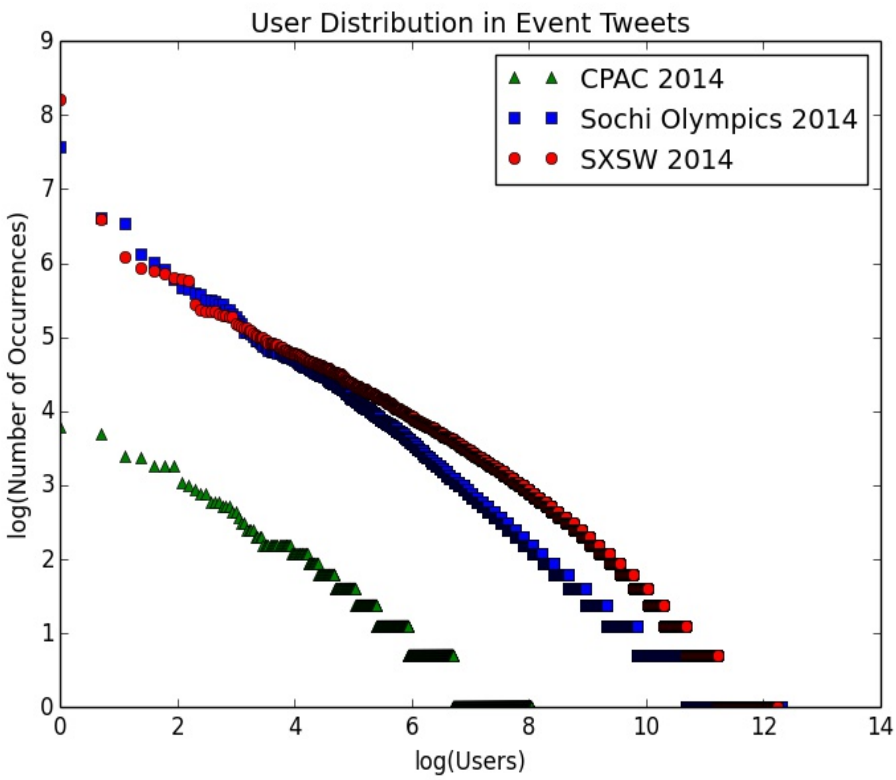


FIGURE 5.9: Distribution of users in event related tweets.



respectively, for the analysis. Some of the characteristics that were prominently different for informative and non-informative tweets are listed in Table 5.10.

FIGURE 5.10: 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%

As presented in the table, some of the observations for all the three events are,

- 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.
- 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 pointed out in Chapter 4, Section 4.2.
- 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. We made the following conclusions based on the observations:

- 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.

5.5.2 Event Identity Information Units

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 [126]. 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. Based on the above analysis we decided to build the EIIS for an event E_i, composed of the following event identity information units:

1. A set of tweets $M_{E_i} = \{m1_{E_i}, m2_{E_i}, ..., mn_{E_i}\}$, related to the event E_i, having high chances of containing informative content.
2. A set of hashtags $H_{E_i} = \{h1_{E_i}, h2_{E_i}, ..., hp_{E_i}\}$, used for annotating the tweets ($\in M_{E_i}$) related to event E_i.
3. A set of text units $W_{E_i} = \{w1_{E_i}, w2_{E_i}, ..., wr_{E_i}\}$, used for expressing textual content in tweets ($\in M_{E_i}$), related to event E_i.
4. A set of URLs $L_{E_i} = \{l1_{E_i}, l2_{E_i}, ..., lt_{E_i}\}$, shared in the tweets ($\in M_{E_i}$) related to event E_i.
5. A set of users $U_{E_i} = \{u1_{E_i}, u2_{E_i}, ..., us_{E_i}\}$, tweeting the tweets ($\in M_{E_i}$), about the event E_i.

5.5.3 Extracting Event Identity Information Units

The event identity information units for an event E_i, as defined above are extracted from the event dataset. Following steps are taken:

- A threshold for the informativeness score assigned in the previous component is set between 0.0-1.0, for extracting the tweets ($\in M_{E_i}$). We set a threshold of 0.7.

Therefore, the tweets having an informativeness score greater than or equal to 0.7 are filtered out and comprises the set M_{E_i}.

- The hashtags that were extracted in the 'Event Reference Preparation' step from the tweets $\in M_{E_i}$, are used for populating the set H_{E_i}. However, the hashtags that matches the slang words and stop words are not considered. This is done in order to ensure good quality of information contextualized by the hashtags.
- The nouns that were extracted in the 'Event Reference Preparation' step from the tweets $\in M_{E_i}$, are considered as the text units ($\in W_{E_i}$). The nouns that matches the slang words are not considered. This is done in order to ensure good quality of textual content represented by the nouns. In another experiment, we considered the extracted named entities as the text units. We report our results and compare the results obtained in both the cases in the next Chapter.
- The expanded URLs from the meta-data of the tweets $\in M_{E_i}$ populates the set L_{E_i}.
- The meta-information of the users posting the tweets $\in M_{E_i}$, represented by their user ids, is extracted for populating the set U_{E_i}

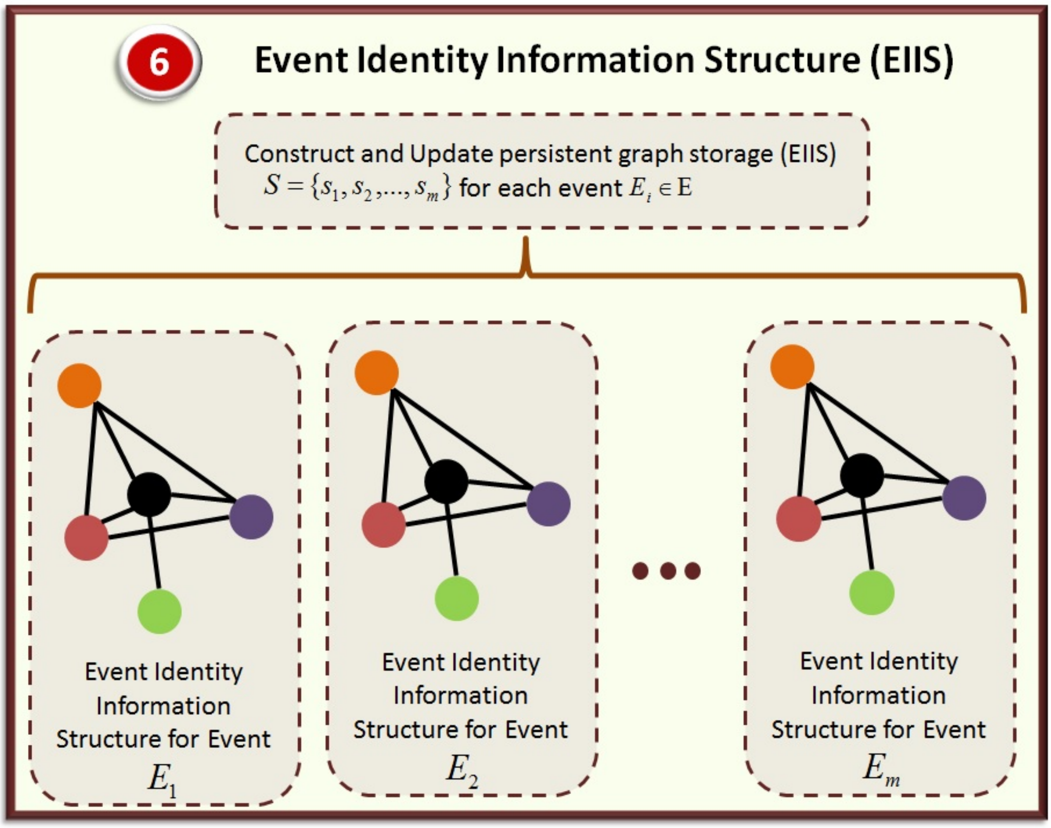
These event identity information units forms the Event Identity Information Structure (EIIS), as explained next.

5.6 Event Identity Information Structure (EIIS)

This is the component that maintains a persistent EIIS as introduced in Chapter 2, Section 2.6, for each individual event tracked by the framework and updates the metadata of the EIIS throughout the EIIM life cycle. Due to the unstructured nature of the social media references and evolving nature of the events, we store the event identity information units extracted by the previous component along with their associated meta-data, in a persistent graph data structure stored in the database. We update the meta-data related to each node of the graph using the normal database updation queries. Adjacency lists are used for representing the graph. The choice of storing the event identity information units in a graph structure is also motivated by the wide array of graph processing algorithms used for natural language processing and text mining operations. We show the efficacy and the advantages of a graph in the next section.

- Therefore the EIIS is a graph $\mathbf{G_{E_i}} = (\mathbf{V_{E_i}}, \mathbf{D_{E_i}})$, where $\mathbf{V_{E_i}} = \mathbf{M_{E_i}} \cup \mathbf{H_{E_i}} \cup \mathbf{W_{E_i}} \cup \mathbf{U_{E_i}} \cup \mathbf{L_{E_i}}$, is the set of vertices and $\mathbf{D_{E_i}}$ is the set of directed edges between different vertices. Whenever two vertices are associated, there are two edges between them that are oppositely directed. For example, if a tweets consist of hashtags, text units, URLs and is posted by an user, then there are bi-directed edges between each one of them.

FIGURE 5.11: Event Identity Information Structure component of the EIIM life cycle.

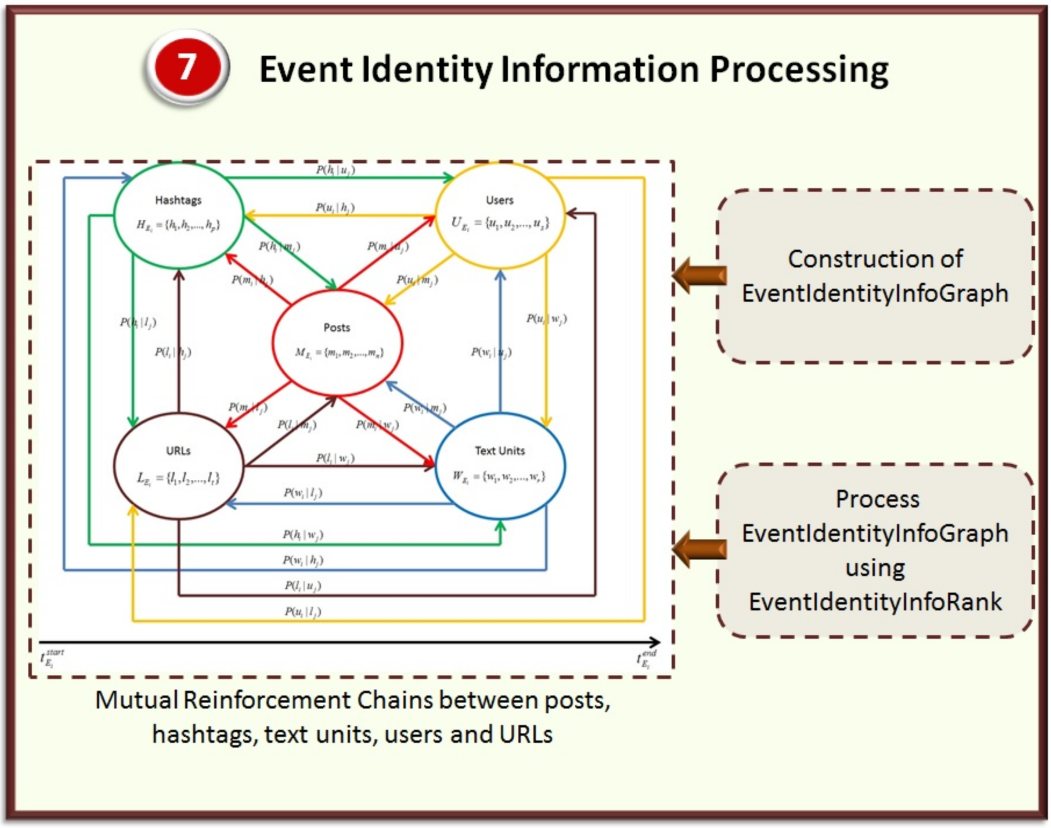


5.7 Event Identity Information Processing

This is the most important processing component of the EIIM life cycle and is at the heart of the entire framework. We make our most novel contributions in this component. The component is mainly divided into two sub-components:

- **EventIdentityInfoGraph** - that represents and defines novel relationships between the vertices of the graph **G** representing the EIIS.
- **EventIdentityInfoProcess** - processes the *EventIdentityInfoGraph* in order to rank its nodes and identify the top most informative event identity information units that acts as inputs to the next two components of the EIIM life cycle.

FIGURE 5.12: Event Identity Information Processing component of the EIIM life cycle.



5.7.1 EventIdentityInfoGraph

We implement a novel graph structure - *EventIdentityInfoGraph*, which is dynamically generated from the graph **G** (EIIS), after a configurable interval of time, using following assumptions.

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,*

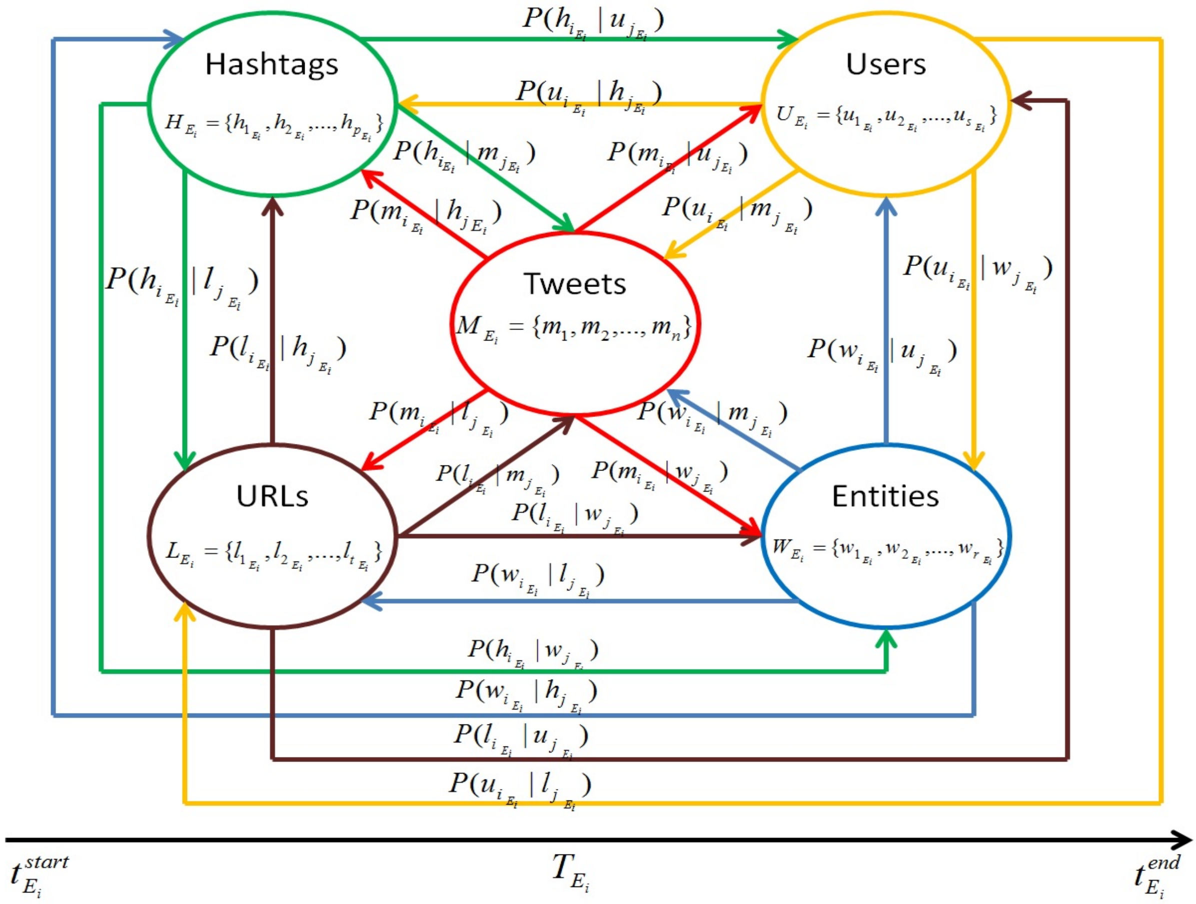
  (d) *event-specific informative users.*

The relationships for an event E_i as stated above, forms a *Mutual Reinforcement Chain* [128] for the event E_i as shown in Figure 5.13. We represent this relationship in a graph $\mathbf{G_{E_i(t)}} = (\mathbf{V_{E_i(t)}}, \mathbf{D_{E_i(t)}})$, which we call as *EventIdentityInfoGraph*, where $V_{E_i(t)} = M_{E_i} \cup H_{E_i} \cup W_{E_i} \cup U_{E_i} \cup L_{E_i}$, is the set of vertices and $\mathbf{D_{E_i(t)}}$ is the set of directed edges between different vertices. The graph $G_{E_i(t)}$ is basically, a snapshot of the EIIS structure of the event E_i at time t.

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 as given by equations 5.1-5.18.

$$P(h_{i_{E_i}} \mid w_{j_{E_i}}) = \frac{No.\,of\,tweets\,h_{i_{E_i}}\,and\,w_{j_{E_i}}\,occur\,together}{No.\,of\,tweets\,w_{j_{E_i}}\,occurs} \tag{5.1}$$

Figure 5.13: Mutual Reinforcement Chains in Twitter for an event.



$$P(w_{i_{E_i}} \mid h_{j_{E_i}}) = \frac{No.\, of\, tweets\, w_{i_{E_i}}\, and\, h_{j_{E_i}}\, occur\, together}{No.\, of\, tweets\, h_{j_{E_i}}\, occurs} \tag{5.2}$$

$$P(h_{i_{E_i}} \mid l_{j_{E_i}}) = \frac{No.\, of\, tweets\, h_{i_{E_i}}\, and\, l_{j_{E_i}}\, occur\, together}{No.\, of\, tweets\, l_{j_{E_i}}\, occurs} \tag{5.3}$$

$$P(l_{i_{E_i}} \mid h_{j_{E_i}}) = \frac{No.\, of\, tweets\, l_{i_{E_i}}\, and\, h_{j_{E_i}}\, occur\, together}{No.\, of\, tweets\, h_{j_{E_i}}\, occurs} \tag{5.4}$$

$$P(h_{i_{E_i}} \mid u_{j_{E_i}}) = \frac{No.\, of\, tweets\, h_{i_{E_i}}\, and\, u_{j_{E_i}}\, occur\, together}{No.\, of\, tweets\, u_{j_{E_i}}\, occurs} \tag{5.5}$$

$$P(u_{i_{E_i}} \mid h_{j_{E_i}}) = \frac{No.\, of\, tweets\, u_{i_{E_i}}\, and\, h_{j_{E_i}}\, occur\, together}{No.\, of\, tweets\, h_{j_{E_i}}\, occurs} \tag{5.6}$$

$$P(w_{i_{E_i}} \mid l_{j_{E_i}}) = \frac{No.\,of\,tweets\,w_{i_{E_i}}\,and\,l_{j_{E_i}}\,occur\,together}{No.\,of\,tweets\,l_{j_{E_i}}\,occurs} \tag{5.7}$$

$$(l_{i_{E_i}} \mid w_{j_{E_i}}) = \frac{No.\,of\,tweets\,l_{i_{E_i}}\,and\,w_{j_{E_i}}\,occur\,together}{No.\,of\,tweets\,w_{j_{E_i}}\,occurs} \tag{5.8}$$

$$P(u_{i_{E_i}} \mid l_{j_{E_i}}) = \frac{No.\,of\,tweets\,u_{i_{E_i}}\,and\,l_{j_{E_i}}\,occur\,together}{No.\,of\,tweets\,l_{j_{E_i}}\,occurs} \tag{5.9}$$

$$P(l_{i_{E_i}} \mid u_{j_{E_i}}) = \frac{No.\,of\,tweets\,l_{i_{E_i}}\,and\,u_{j_{E_i}}\,occur\,together}{No.\,of\,tweets\,u_{j_{E_i}}\,occurs} \tag{5.10}$$

$$P(h_{i_{E_i}} \mid m_{j_{E_i}}) = 1.0 \tag{5.11}$$

$$P(m_{i_{E_i}} \mid h_{j_{E_i}}) = 1.0 \tag{5.12}$$

$$P(w_{i_{E_i}} \mid m_{j_{E_i}}) = 1.0 \tag{5.13}$$

$$P(m_{i_{E_i}} \mid w_{j_{E_i}}) = 1.0 \tag{5.14}$$

$$P(u_{i_{E_i}} \mid m_{j_{E_i}}) = 1.0 \tag{5.15}$$

$$P(m_{i_{E_i}} \mid u_{j_{E_i}}) = 1.0 \tag{5.16}$$

$$P(l_{i_{E_i}} \mid m_{j_{E_i}}) = 1.0 \tag{5.17}$$

$$P(m_{i_{E_i}} \mid l_{j_{E_i}}) = 1.0 \tag{5.18}$$

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 as explained in Chapter 4, Section 4.2.

Next, we explain *EventIdentityInfoRank*.

5.7.2 EventIdentityInfoRank

EventIdentityInfoRank is an iterative algorithm that takes into account the mutually reinforcing relationships between the vertices of *EventIdentityInfoGraph* as explained in the previous section and propagates event-specific scores of each vertex to connected vertices across the graph for ranking its vertices ($\in V_{E_i(t)}$) 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 5.19-5.22. The tweets ($\in M_{E_i}$) are assigned an initial informativeness score as obtained from the logistic regression model in 'Event Information Quality' component. 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 *EventIdentityInfoGraph*. We aim to refine the initial scores and assign a final score for ranking the vertices by leveraging the mutually reinforcing relationships between them.

$$Score(h_{i_{E_i}}) = \frac{freq(h_i)}{max\{freq(h_1), freq(h_2), ..., freq(h_p)\}} \tag{5.19}$$

$$Score(w_{i_{E_i}}) = \frac{freq(w_i)}{max\{freq(w_1), freq(w_2), ..., freq(w_r)\}} \tag{5.20}$$

$$Score(u_{i_{E_i}}) = \frac{followers(u_i)}{max\{followers(u_1), ..., followers(u_r)\}} \tag{5.21}$$

$$Score(l_{i_{E_i}}) = \frac{freq(l_i)}{max\{freq(l_1), freq(l_2), ..., freq(l_r)\}} \tag{5.22}$$

The relationships between two different subsets of vertices in graph $\mathbf{G_{E_i(t)}}$ 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 $(\mathbf{i}, \mathbf{j})^{\mathbf{th}}$ 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 equations 5.1-5.18. 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:

$$R^{M(k+1)}_{E_i} = A^{MM(k)}_{E_i} R^{M(k)}_{E_i} + A^{MH(k)}_{E_i} R^{H(k)}_{E_i} + A^{MW(k)}_{E_i} R^{W(k)}_{E_i} + A^{MU(k)}_{E_i} R^{U(k)}_{E_i} + A^{ML(k)}_{E_i} R^{L(k)}_{E_i} \tag{5.23}$$

$$R^{H(k+1)}_{E_i} = A^{HM(k)}_{E_i} R^{M(k)}_{E_i} + A^{HH(k)}_{E_i} R^{H(k)}_{E_i} + A^{HW(k)}_{E_i} R^{W(k)}_{E_i} + A^{HU(k)}_{E_i} R^{U(k)}_{E_i} + A^{HL(k)}_{E_i} R^{L(k)}_{E_i} \tag{5.24}$$

$$R^{W(k+1)}_{E_i} = A^{WM(k)}_{E_i} R^{M(k)}_{E_i} + A^{WH(k)}_{E_i} R^{H(k)}_{E_i} + A^{WW(k)}_{E_i} R^{W(k)}_{E_i} + A^{WU(k)}_{E_i} R^{U(k)}_{E_i} + A^{WL(k)}_{E_i} R^{L(k)}_{E_i} \tag{5.25}$$

$$R^{U(k+1)}_{E_i} = A^{UM(k)}_{E_i} R^{M(k)}_{E_i} + A^{UH(k)}_{E_i} R^{H(k)}_{E_i} + A^{UW(k)}_{E_i} R^{W(k)}_{E_i} + A^{UU(k)}_{E_i} R^{U(k)}_{E_i} + A^{UL(k)}_{E_i} R^{L(k)}_{E_i} \tag{5.26}$$

$$R^{L(k+1)}_{E_i} = A^{LM(k)}_{E_i} R^{M(k)}_{E_i} + A^{LH(k)}_{E_i} R^{H(k)}_{E_i} + A^{LW(k)}_{E_i} R^{W(k)}_{E_i} + A^{LU(k)}_{E_i} R^{U(k)}_{E_i} + A^{LL(k)}_{E_i} R^{L(k)}_{E_i} \tag{5.27}$$

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

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

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

$$\Delta_{E_i}.R_{E_i} = \lambda.R_{E_i} \tag{5.28}$$

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

$$\overline{\Delta}_{E_i} = \alpha\hat{\Delta}_{E_i} + (1-\alpha)E \tag{5.29}$$

$$E = p \times [1]_{1\times k} \tag{5.30}$$

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 $\mathbf{\hat{M}_{E_i}}, \mathbf{\hat{H}_{E_i}}, \mathbf{\hat{W}_{E_i}}, \mathbf{\hat{L}_{E_i}}, \mathbf{\hat{U}_{E_i}}$, thus obtained are the tweets, hashtags, text units, URLs and users, ranked in terms of their event-specific informativeness. The entire procedure is presented step-by-step in an Algorithm 1.

Since the algorithm uses power iteration method for ranking the vertices of a graph, it could be easily made scalable using mapreduce paradigm [129]. We plan to work on it in the future and implement our framework using hadoop and mapreduce environment. Also, the EIIM 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.

5.8 Event Reference Resolution

The output of the previous component assigns a final ranked event-specific informativeness scores to the hashtags as well as the text units. These scores are specific to an event and are different for the same text unit or hashtag when they appear in EIIS of some other event. A threshold is set for filtering out the hashtags and text units that

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

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

Algorithm 1: EventIdentityInfoRank algorithm

are highly ranked. These hashtags and text units are used for assigning scores to the incoming tweets after tokenizing them to its constituent text units and hashtags. Another threshold is set to decide the acceptable score of a tweet in order to consider it as a high quality informative tweet related to an event tracked by the framework. In this way the incoming real-time tweets are resolved against a particular event, which are stored with a reference to the respective event identifier. These tweets are further processed in the EIIM life cycle.

5.9 Event Analytics

Top Five Event-specific Informative Hashtags for Sydney Siege Event

1. #sydneysiege
2. #SydneySiege
3. #Sydneysiege
4. #MartinPlace

Figure 5.14: Event Reference Resolution component of the EIIM life cycle.

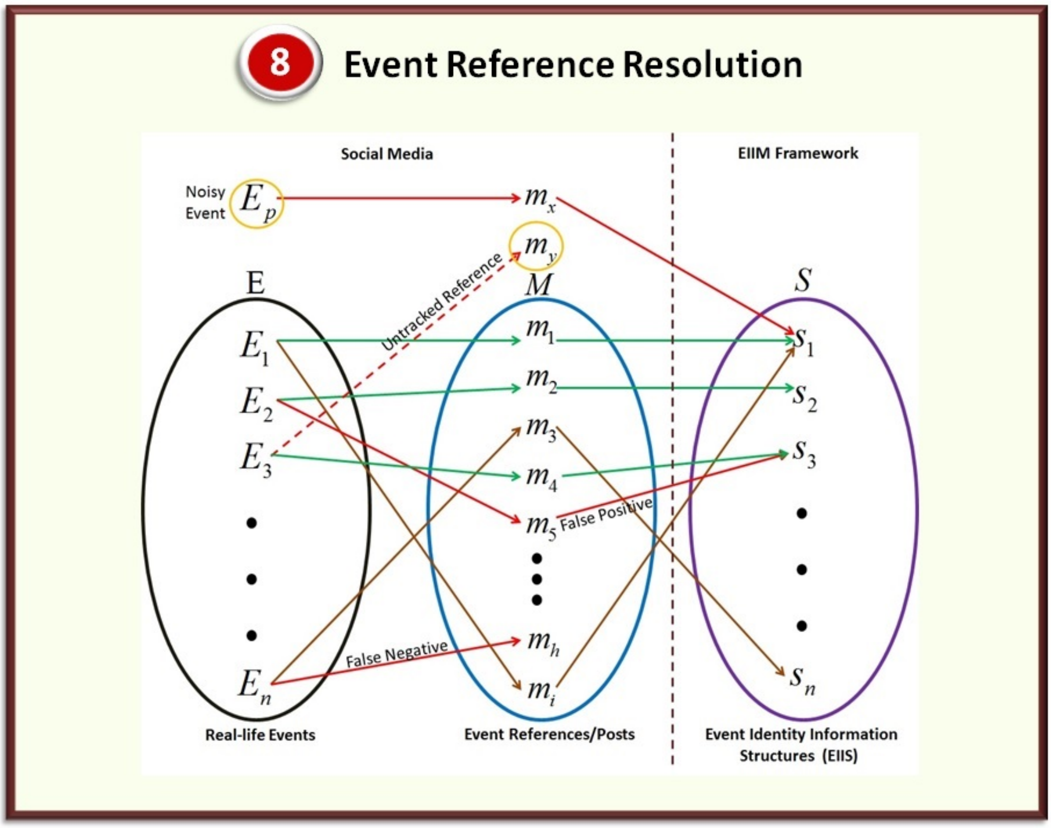
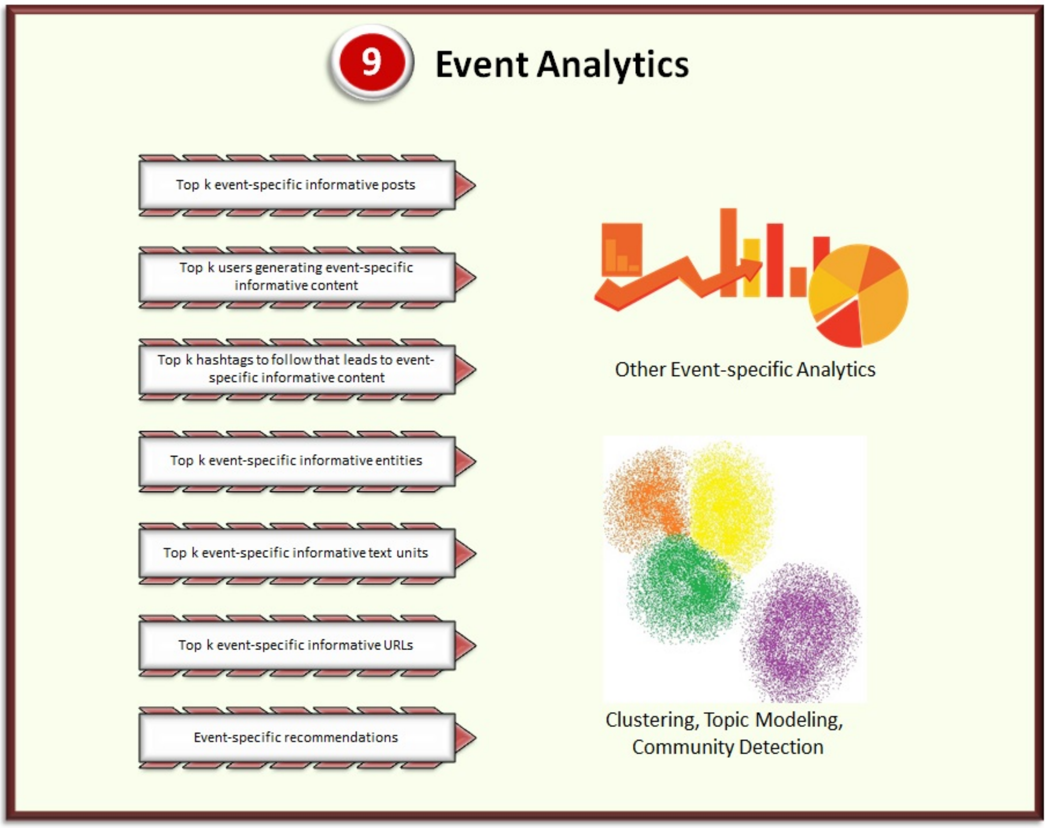


Figure 5.15: Event Analytics component of the EIIM life cycle.



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...

5.10 Event Identity Information Processing

Chapter 6

Evaluations

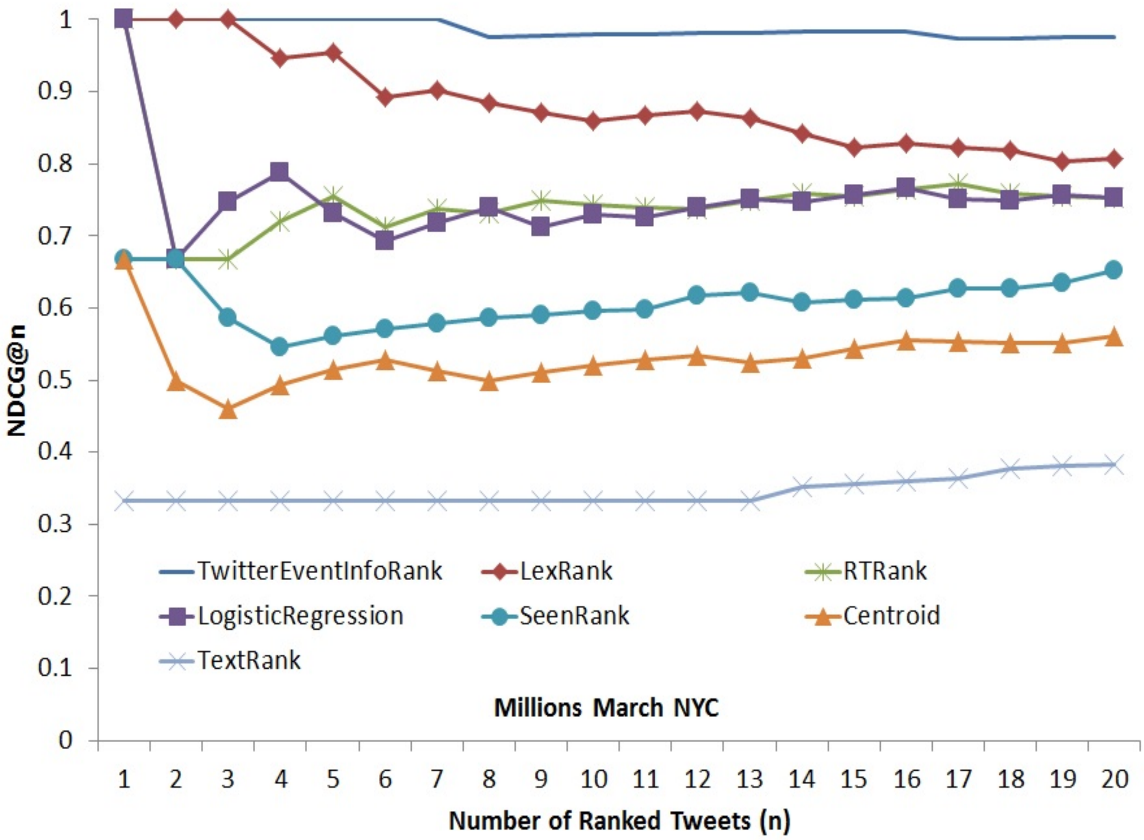


FIGURE 6.1: Performance comparison of ranking techniques using NDCG scores.

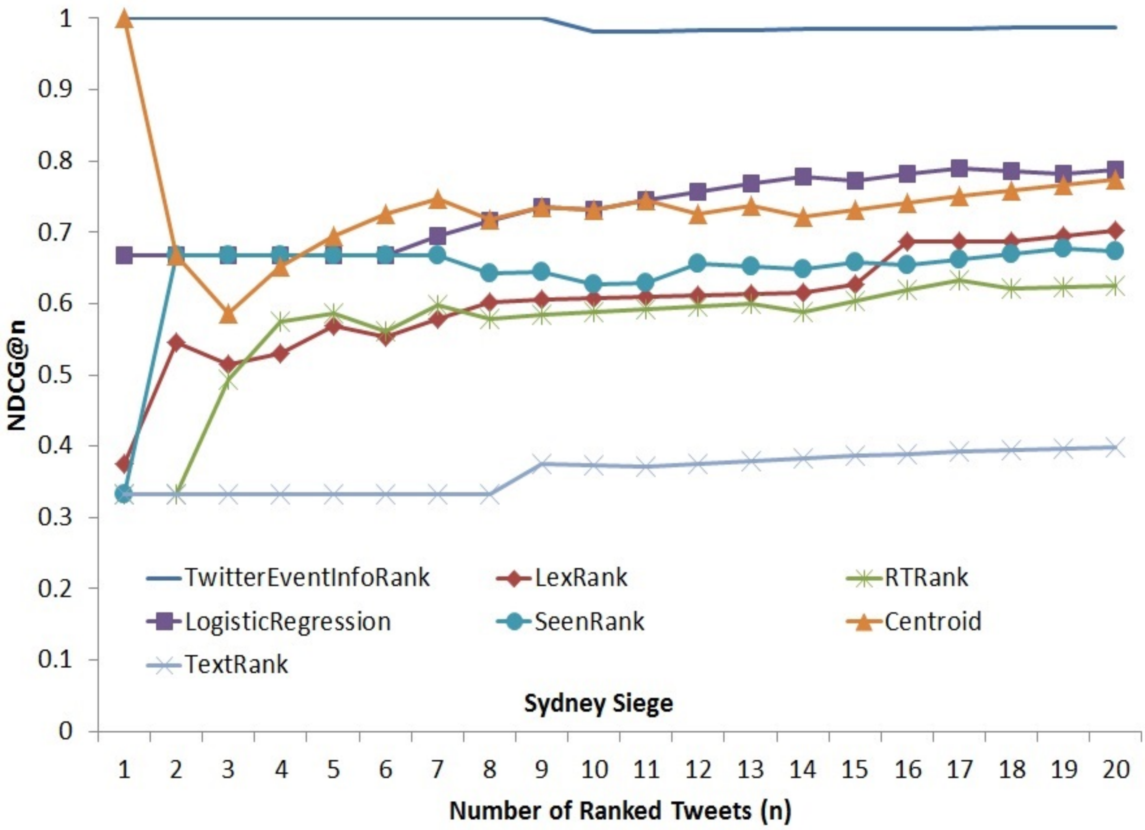


FIGURE 6.2: Performance comparison of ranking techniques using NDCG scores.

TABLE 6.1: 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

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 6.3: 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 6.4: 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 6.5: 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 6.6: Performance comparison of ranking techniques using precision scores.

TABLE 6.2: 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

Chapter 7

Potential Applications of the EIIM Framework

7.1 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' [130] to international sporting events like the 'Winter Olympics' [131] to socio-political [5] and socio-economical [132] events that shook the world such as presidential elections [21], 'Egyptian Revolution' [133], 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 [134]. Social media is being used for tracking terrorism activities [135], collective actions [13], 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/

7.2 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.

7.3 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.

7.4 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.

7.5 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.

7.6 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 8

Conclusion and Future Work

8.1 Conclusion

8.2 Future Work

8.2.1 Summarizing Event Related Content

8.2.2 Identifying Insightful Opinionated Content Related to Events

8.2.3 Event Topic Modeling

8.2.4 Event-specific Recommendations

8.2.5 Distributed Processing of EventIdentityInfoGraph

8.2.6 Event Ontology for Social Media

Appendix A

Appendix Title Here

Write your Appendix content here.

Bibliography

[1] Yinle Zhou and John Talburt. Entity identity information management (eiim). In *International Conference on Information Quality (ICIQ-11), Adelaide, Australia*, pages 327–341, 2011.

[2] 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.

[3] 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.

[4] Vivek Kumar Singh, Rakesh Adhikari, and Debanjan Mahata. A clustering and opinion mining approach to socio-political analysis of the blogosphere. In *Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on*, pages 1–4. IEEE, 2010.

[5] 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.

[6] 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.

[7] 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.

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

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

[10] Nitin Agarwal and Debanjan Mahata. Grouping the similar among the disconnected bloggers. *Social Media Mining and Social Network Analysis: Emerging Research: Emerging Research*, page 54, 2013.

[11] Nitin Agarwal, Debanjan Mahata, and Huan Liu. Time-and event-driven modeling of blogger influence. In *Encyclopedia of Social Network Analysis and Mining*, pages 2154–2165. Springer, 2014.

[12] Fatih Sen, Rolf T Wigand, Nitin Agarwal, Debanjan Mahata, and Halil Bisgin. Identifying focal patterns in social networks. In *CASoN*, pages 105–108, 2012.

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

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

[15] 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.

[16] 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.

[17] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.

[18] Nitin Agarwal and Yusuf Yiliyasi. Information quality challenges in social media. In *International Conference on Information Quality (ICIQ 2010), Little Rock, Arkansas*, 2010.

[19] Naila Hamdy and Ehab H Gomaa. Framing the egyptian uprising in arabic language newspapers and social media. *Journal of Communication*, 62(2):195–211, 2012.

[20] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.

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

[22] Social media for events, 2014. URL http://socialmediaforeventsebook.com/.

[23] Nina Eyrich, Monica L Padman, and Kaye D Sweetser. Pr practitioners' use of social media tools and communication technology. *Public relations review*, 34(4): 412–414, 2008.

[24] W Glynn Mangold and David J Faulds. Social media: The new hybrid element of the promotion mix. *Business horizons*, 52(4):357–365, 2009.

[25] Sasa Petrovic, Miles Osborne, Richard McCreadie, Craig Macdonald, and Iadh Ounis. Can twitter replace newswire for breaking news? 2013.

[26] How bin laden news spread on twitter, 2011. URL http://www.cnn.com/2011/TECH/social.media/05/02/osama.bin.laden.twitter/.

[27] Theater shooting unfolds in real time on social media, 2012. URL http://www.cnn.com/2012/07/20/tech/social-media/colorado-shooting-social-media/.

[28] Google and twitter launch service enabling egyptians to tweet by phone, 2011. URL http://www.theguardian.com/technology/2011/feb/01/google-twitter-egypt.

[29] John R Talburt. *Entity resolution and information quality*. Elsevier, 2011.

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

[31] 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.

[32] Sidharth Chhabra, Anupama Aggarwal, Fabricio Benevenuto, and Ponnurangam Kumaraguru. Phi. sh/$ ocial: the phishing landscape through short urls. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, pages 92–101. ACM, 2011.

[33] Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 27–37. ACM, 2010.

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

[35] Emma Tonkin, Heather D Pfeiffer, and Greg Tourte. Twitter, information sharing and the london riots? *Bulletin of the American Society for Information Science and Technology*, 38(2):49–57, 2012.

[36] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[37] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 183–194. ACM, 2008.

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

[39] Aditi Gupta and Ponnurangam Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, page 2. ACM, 2012.

[40] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, pages 249–252. ACM, 2011.

[41] Firas Damak, Karen Pinel-Sauvagnat, Mohand Boughanem, and Guillaume Cabanac. Effectiveness of state-of-the-art features for microblog search. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 914–919. ACM, 2013.

[42] 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.

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

[44] 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.

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

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

[47] Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 295–303. Association for Computational Linguistics, 2010.

[48] 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.

[49] 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.

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

[51] David Inouye and Jugal K Kalita. Comparing twitter summarization algorithms for multiple post summaries. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 298–306. IEEE, 2011.

[52] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 189–198. ACM, 2012.

[53] Beaux Sharifi, M-A Hutton, and Jugal K Kalita. Experiments in microblog summarization. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 49–56. IEEE, 2010.

[54] Günes Erkan and Dragomir R Radev. Lexrank: graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, pages 457–479, 2004.

[55] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. Association for Computational Linguistics, 2004.

[56] Ivan P Fellegi and Alan B Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[57] Howard B Newcombe, James M Kennedy, SJ Axford, and Allison P James. Automatic linkage of vital records computers can be used to extract" follow-up" statistics of families from files of routine records. *Science*, 130(3381):954–959, 1959.

[58] Thomas N Herzog, Fritz J Scheuren, and William E Winkler. *Data quality and record linkage techniques*. Springer Science & Business Media, 2007.

[59] Mauricio A Hernández and Salvatore J Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data mining and knowledge discovery*, 2 (1):9–37, 1998.

[60] Sunita Sarawagi and Anuradha Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–278. ACM, 2002.

[61] Rohit Ananthakrishna, Surajit Chaudhuri, and Venkatesh Ganti. Eliminating fuzzy duplicates in data warehouses. In *Proceedings of the 28th international conference on Very Large Data Bases*, pages 586–597. VLDB Endowment, 2002.

[62] Y Richard Wang and Stuart E Madnick. The inter-database instance identification problem in integrating autonomous systems. In *Data Engineering, 1989. Proceedings. Fifth International Conference on*, pages 46–55. IEEE, 1989.

[63] William W Cohen, Henry Kautz, and David McAllester. Hardening soft information sources. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 255–259. ACM, 2000.

[64] Mikhail Bilenko, Raymond Mooney, William Cohen, Pradeep Ravikumar, and Stephen Fienberg. Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23, 2003.

[65] Ahmed K Elmagarmid, Panagiotis G Ipeirotis, and Vassilios S Verykios. Duplicate record detection: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 19(1):1–16, 2007.

[66] Hector Garcia-Molina. Pair-wise entity resolution: overview and challenges. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 1–1. ACM, 2006.

[67] Omar Benjelloun, Hector Garcia-Molina, Hideki Kawai, Tait Eliott Larson, David Menestrina, Qi Su, Sutthipong Thavisomboon, and Jennifer Widom. Generic entity resolution in the serf project. *IEEE Data Engineering Bulletin, June 2006 Issue*, 2006.

[68] Omar Benjelloun, Hector Garcia-Molina, Heng Gong, Hideki Kawai, Tait Eliott Larson, David Menestrina, and Sutthipong Thavisomboon. D-swoosh: A family of algorithms for generic, distributed entity resolution. In *Distributed Computing Systems, 2007. ICDCS'07. 27th International Conference on*, pages 37–37. IEEE, 2007.

[69] John Talburt, Richard Wang, Kimberly Hess, and Emily Kuo. An algebraic approach to data quality metrics for entity resolution over large datasets. *Information quality management: Theory and applications*, pages 1–22, 2007.

[70] Jenny Rose Finkel. Named entity recognition and the stanford ner software, 2007.

[71] Jason Baldridge. The opennlp project. *URL: http://opennlp. apache. org/index. html,(accessed 2 February 2012)*, 2005.

[72] Hamish Cunningham. Gate, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254, 2002.

[73] Breck Baldwin and Bob Carpenter. Lingpipe. *Available from World Wide Web: http://alias-i. com/lingpipe*, 2003.

[74] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

[75] Jakub Piskorski and Roman Yangarber. Information extraction: Past, present and future. In *Multi-source, multilingual information extraction and summarization*, pages 23–49. Springer, 2013.

[76] Sunita Sarawagi. Information extraction. *Foundations and trends in databases*, 1 (3):261–377, 2008.

[77] Wen Hua, Dat T Huynh, Saeid Hosseini, Jiaheng Lu, and Xiaofang Zhou. Information extraction from microblogs: A survey. *Int. J. Soft. and Informatics*, 6(4): 495–522, 2012.

[78] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[79] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM, 2008.

[80] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. *Dbpedia: A nucleus for a web of open data.* Springer, 2007.

[81] Omkar Deshpande, Digvijay S Lamba, Michel Tourn, Sanjib Das, Sri Subramaniam, Anand Rajaraman, Venky Harinarayan, and AnHai Doan. Building, maintaining, and using knowledge bases: A report from the trenches. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 1209–1220. ACM, 2013.

[82] Abhishek Gattani, Digvijay S Lamba, Nikesh Garera, Mitul Tiwari, Xiaoyong Chai, Sanjib Das, Sri Subramaniam, Anand Rajaraman, Venky Harinarayan, and AnHai Doan. Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach. *Proceedings of the VLDB Endowment*, 6(11): 1126–1137, 2013.

[83] Mustafa Bilgic, Louis Licamele, Lise Getoor, and Ben Shneiderman. D-dupe: An interactive tool for entity resolution in social networks. In *Visual Analytics Science And Technology, 2006 IEEE Symposium On*, pages 43–50. IEEE, 2006.

[84] Lars Kolb, Andreas Thor, and Erhard Rahm. Dedoop: efficient deduplication with hadoop. *Proceedings of the VLDB Endowment*, 5(12):1878–1881, 2012.

[85] John R Talburt and Yinle Zhou. *Entity Information Life Cycle for Big Data: Master Data Management and Information Integration.* Morgan Kaufmann, 2015.

[86] Paolo Bouquet and Stefano Bortoli. Entity-centric social profile integration. In *Proceedings of the International Workshop on Linking of User Profiles and Applications in the Social Semantic Web (LUPAS 2010)*, pages 52–57, 2010.

[87] Stefano Bortoli, Heiko Stoermer, Paolo Bouquet, and Holger Wache. Foaf-o-matic-solving the identity problem in the foaf network. In *SWAP*, 2007.

[88] Elie Raad, Richard Chbeir, and Albert Dipanda. User profile matching in social networks. In *Network-Based Information Systems (NBiS), 2010 13th International Conference on*, pages 297–304. IEEE, 2010.

[89] Paolo Bouquet, Heiko Stoermer, Michele Mancioppi, and Daniel Giacomuzzi. Okkam: Towards a solution to the"identity crisis"on the semantic web. In *SWAP*, volume 201, 2006.

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

[91] 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.

[92] 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.

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

[94] 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.

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

[96] 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.

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

[98] 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.

[99] 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.

[100] 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.

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

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] Hila Becker, Mor Naaman, and Luis Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 291–300. ACM, 2010.

[107] 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.

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

[109] 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.

[110] Hila Becker, Dan Iter, Mor Naaman, and Luis Gravano. Identifying content for planned events across social media sites. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 533–542. ACM, 2012.

[111] Hila Becker, Feiyang Chen, Dan Iter, Mor Naaman, and Luis Gravano. Automatic identification and presentation of twitter content for planned events. In *ICWSM*, 2011.

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

[113] Daniel Gayo-Avello. Nepotistic relationships in twitter and their impact on rank prestige algorithms. *Information Processing & Management*, 49(6):1250–1280, 2013.

[114] 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.

[115] 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.

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

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

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

[119] 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.

[120] 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.

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

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

[123] 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.

[124] Minlie Huang, Yi Yang, and Xiaoyan Zhu. Quality-biased ranking of short texts in microblogging services. In *IJCNLP*, pages 373–382, 2011.

[125] Mosquera Alejandro and Moreda Paloma. The use of metrics for measuring informality levels in web 2.0 texts. 2011.

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

[127] 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.

[128] 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.

[129] 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.

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

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

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

[133] 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.

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

[135] 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.