University of Arkansas at Little Rock

Doctoral Thesis

# A Framework for Collecting, Extracting and Managing Event Identity Information from Short Social Media Text

*Author:*
Debanjan Mahata

*Supervisor:*
Dr. John R. Talburt

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

*in*

Integrated Computing
Information Quality Track
Department of Information Science

April 2015

# Declaration of Authorship

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

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

Signed: _____

Date: _____

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

Ronald Coase, Economics, Nobel Prize Laureate

# *Abstract*

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

# *Acknowledgements*

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

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

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

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

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

# Contents

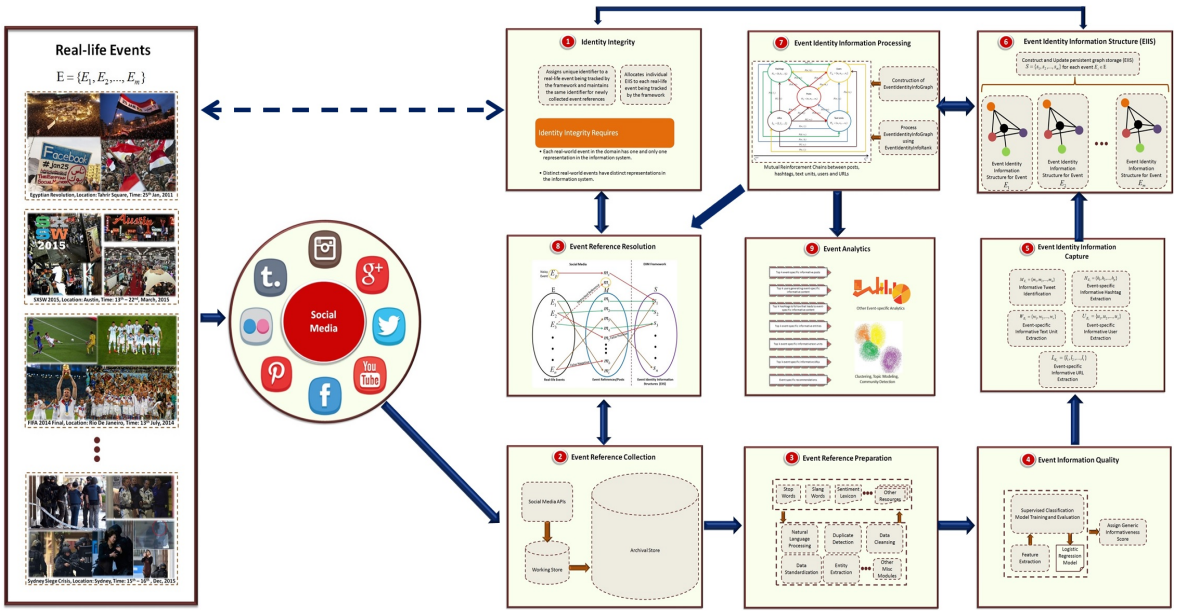
# List of Figures

# List of Tables

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

# Dissertation Overview

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




## Related Publications

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

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

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

- Nitin Agarwal, **Debanjan Mahata**, and Huan Liu. *Time-and Event-Driven Modeling of Blogger Influence*. Encyclopedia of Social Network Analysis and Mining. Springer New York, 2014. 2154-2165.

- **Debanjan Mahata** and Nitin Agarwal. *Learning from the crowd: An Evolutionary Mutual Reinforcement Model for Analyzing Events*. Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. IEEE, 2013.

- Nitin Agarwal, and **Debanjan Mahata**. *Grouping the Similar among the Disconnected Bloggers.* Social Media Mining and Social Network Analysis: Emerging Research (2013), 54.

- **Debanjan Mahata**, and Nitin Agarwal. *What does everybody know? identifying event-specific sources from social media.* IEEE Fourth International Conference on Computational Aspects of Social Networks (CASoN), 2012.

- **Debanjan Mahata** and Nitin Agarwal. *Analyzing Event-specific Socio-Technical Behaviors Through the Lens of Social Media.* The International Sunbelt Social Network Conference (Sunbelt XXXII) organized by the International Network for Social Network Analysis (INSNA), March 12-18, 2012, Redondo Beach, California.

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

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

# Related Submitted Publications

- **Debanjan Mahata**, John R. Talburt, Vivek Kumar Singh and Rajesh Piryani; *Chatter that Matter: A Framework for Identifying and Ranking Event-specific Informative Tweets.* $18^{th}$ International Conference on Text, Speech and Dialogue, Plzen, Czech Republic (Notification Due: May 10, 2015)

- **Debanjan Mahata**, John R. Talburt and Vivek Kumar Singh; *A Framework for Collecting, Extracting and Managing Event Identity Information from Twitter.* $20^{th}$ International Conference on Information Quality, M.I.T, Boston (Notification Due: April 30, 2015)

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

# Chapter 1

# Introduction

## 1.1 Social Media and Real-life Events

## 1.2 Challenges and Opportunities

## 1.3 Research Questions

## 1.4 Research Methodology

## 1.5 Research Contributions

## 1.6 Structure of the Thesis

# Chapter 2

# Literature Review

## 2.1 Event Identification in News Text

## 2.2 Event Identification in Social Media

## 2.3 Information Quality in Social Media

## 2.4 Ranking and Summarization of Short Textual Social Media Posts

## 2.5 Reference Tracking and Entity Resolution

# Chapter 3

# Defining Events

## 3.1 Topic Detection and Tracking

## 3.2 Automatic Content Extraction

## 3.3 Multimedia Event Detection

## 3.4 Events in Social Media

# Chapter 4

# Event Identity Information Management Life Cycle

## 4.1 Identity Integrity

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

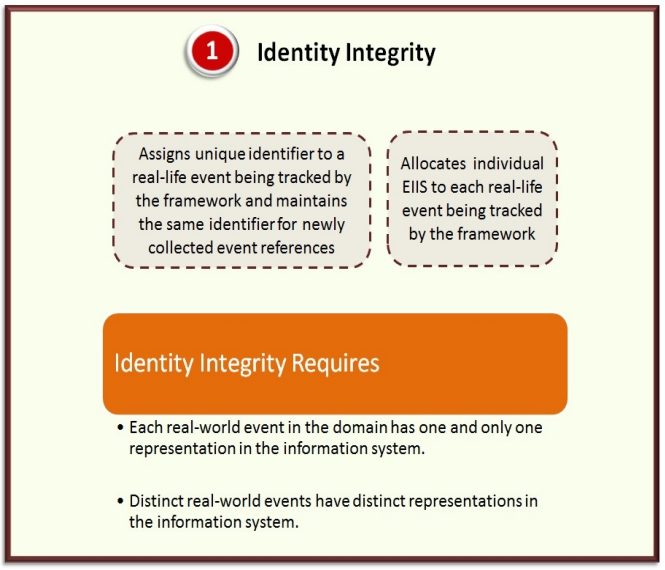


FIGURE 4.2: Event Reference Collection component of the EIIM life cycle.

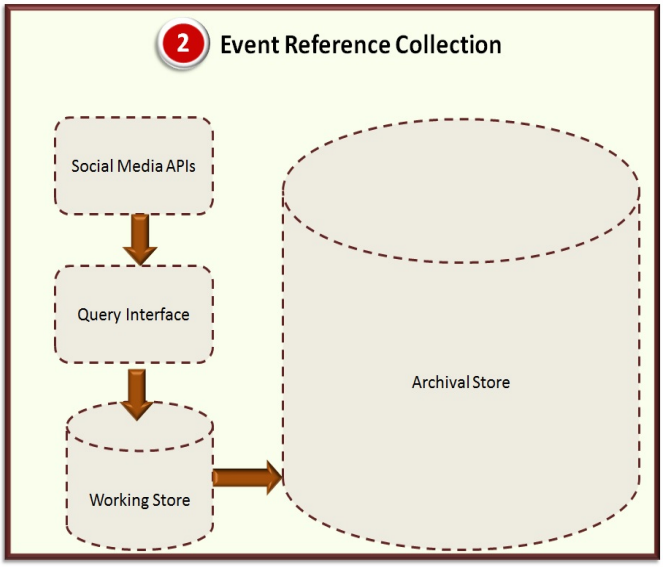



FIGURE 4.3: Event Reference Preparation component of the EIIM life cycle.

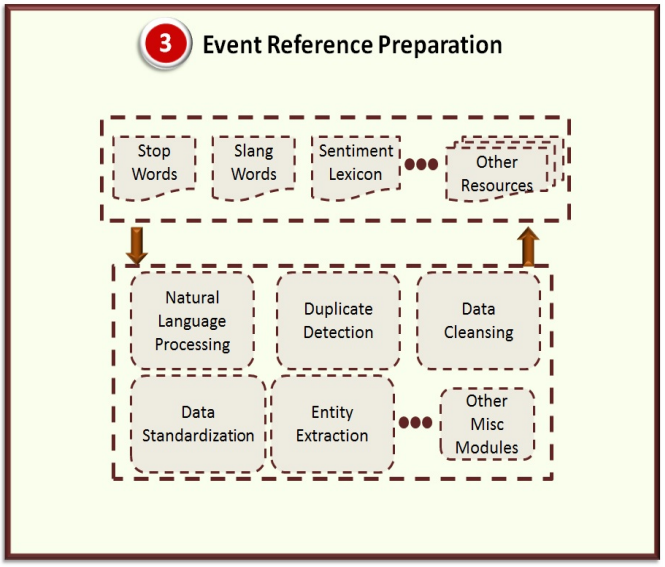


FIGURE 4.4: Event Information Quality component of the EIIM life cycle.

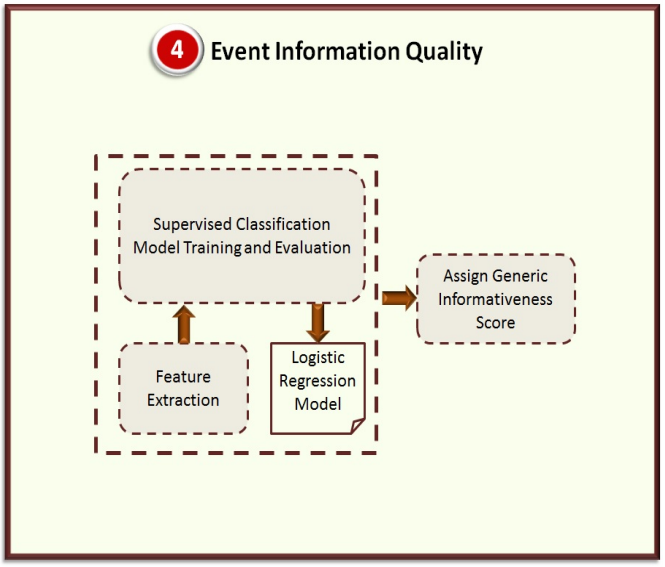
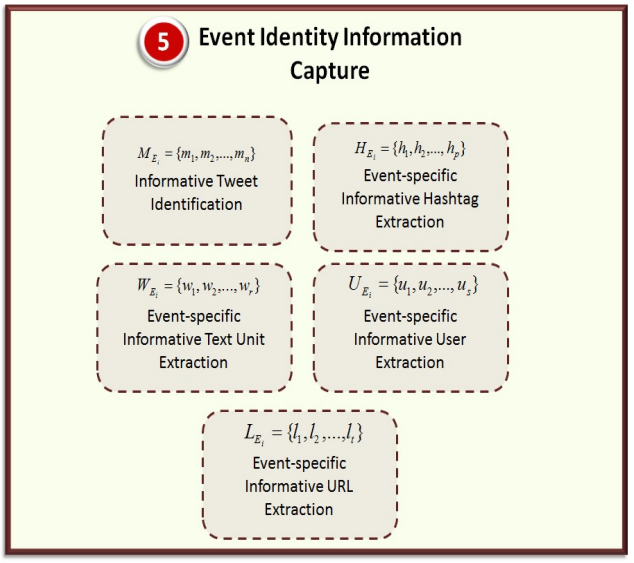



FIGURE 4.5: Event Identity Information Capture component of the EIIM life cycle.




## 4.2 Event Reference Collection

## 4.3 Event Reference Preparation

## 4.4 Event Information Quality

## 4.5 Event Identity Information Capture

## 4.6 Event Identity Information Structure

## 4.7 Event Identity Information Processing

## 4.8 Event Reference Resolution

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

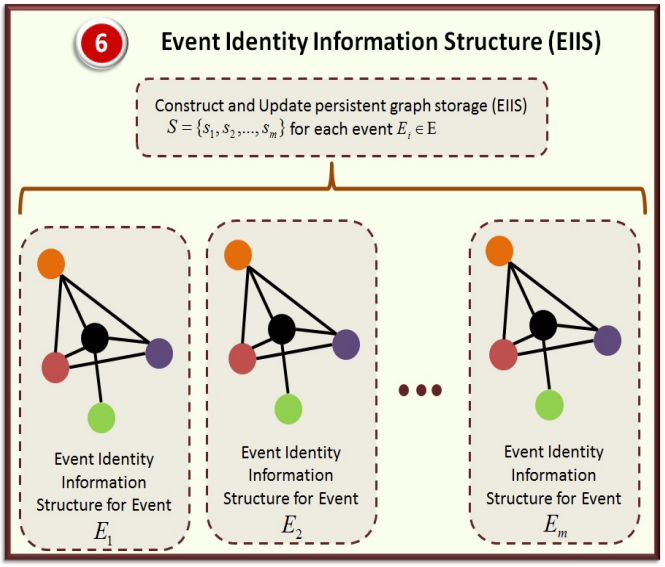



FIGURE 4.7: Event Identity Information Processing component of the EIIM life cycle.

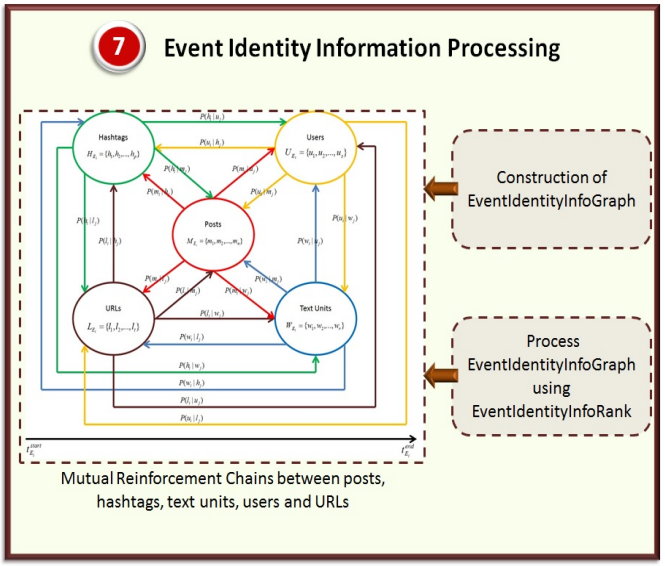


FIGURE 4.8: Event Reference Resolution component of the EIIM life cycle.

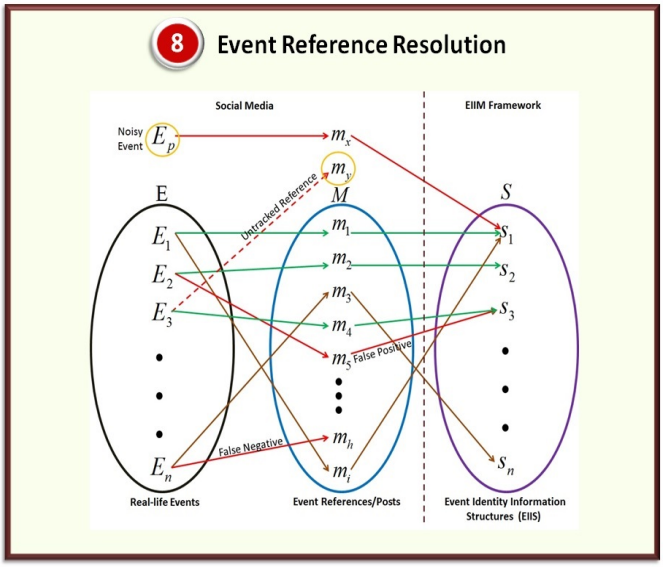



FIGURE 4.9: Event Analytics component of the EIIM life cycle.

# Chapter 5

# Discovering Event-specific Informative Content from Twitter

## 5.1 Twitter and Event Related Content

## 5.2 Analysis of Informative and Non-informative Content in Tweets

## 5.3 EventIdentityInfoGraph

## 5.4 EventIdentityInfoRank

## 5.5 Experiments

### 5.5.1 Data Collection

### 5.5.2 Data Preparation

### 5.5.3 Baselines

### 5.5.4 Evaluation

### 5.5.5 Sample Results

# Chapter 6

# Applications

## 6.1 Event Monitoring and Analysis

References related to real-life events are extremely abundant in social media. Right from natural disasters such as the 'Haiti Earthquake' to international sporting events like the 'Winter Olympics' to socio-political and socio-economical events that shook the world such as presidential elections, 'Egyptian Revolution', 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 [1]. Social media is being used for tracking terrorism activities [2], collective actions [3], and countering cyber-attack threats[6]. One of the main components of each of these applications is tracking references related to the events. The proposed EIIM model could be an essential component of such systems. It would help in identifying, tracking and analyzing events and its related references in an organized manner over time.

[1] http://geofeedia.com/
[2] http://twitterstand.umiacs.umd.edu/
[3] http://twitris.knoesis.org/
[4] http://truthy.indiana.edu/
[5] http://tweettracker.fulton.asu.edu/
[6] https://www.recordedfuture.com/

## 6.2 Event Information Retrieval

## 6.3 Event Opinion Mining

## 6.4 Event-specific Recommendations

## 6.5 Event Management and Marketing

## 6.6 Social Media Data Integration

Organizations have increasingly started integrating the data available in social media with the enterprise data[7]. 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 products. An EIIM system capable of operating in social media could go a long way in collecting the 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.

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

# Chapter 7

# Conclusion and Future Work

## 7.1 Conclusion

## 7.2 Future Work

### 7.2.1 Summarizing Event Related Content

### 7.2.2 Identifying Insightful Opinionated Content Related to Events

### 7.2.3 Event Topic Modeling

### 7.2.4 Event-specific Recommendations

### 7.2.5 Distributed Processing of EventIdentityInfoGraph

### 7.2.6 Event Ontology for Social Media

# Appendix A

# Appendix Title Here

Write your Appendix content here.

# Bibliography

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

[2] 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.

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