**A Framework for Collecting, Extracting and Managing Event Identity Information from Twitter**

Completed Academic Paper

**Abstract:** With the popularity of Twitter, there has been voluminous growth in the digital footprints of real-life events in the Internet. The references to different types of events in Twitter have the potential to provide extremely valuable information to researchers and organizations, which could be mined and analyzed for making major decisions. There are tremendous applications in the areas of real-life event analysis, opinion mining, reference tracking, online advertising, recommendation engines, cyber security, event management, enterprise data integration, among others. Thus, there is a need of a generic framework that can collect different event references, extract identity information of the events from them and maintain the information persistently for resolving new references to the events and provide updated analytics. The presented research establishes the design and implementation of such a framework from the perspective of *Event Identity Information Management* (EIIM) in the domain of Twitter. The paper introduces the problem of EIIM in Twitter, discusses the prevalent challenges and proposes the design of a framework capable of managing persistent identity information of pre-specified set of events. We explore the applications of the research, validate the different components of the framework and conclude with our comments on various criteria showing high efficacy and practical utility of our proposed framework.

**Keywords:** entity resolution, twitter, information quality, event identity information management, event identity information structure, information integration, social media data integration, enterprise data integration, ranking, information retrieval.

**Introduction**

Twitter has become one of the most popular social media platforms for sharing real-time updates and information related to different events occurring across the world (e.g. football matches, conferences, music festivals, protest movements, etc). Studies have shown it to be a medium for circulating trending news through citizen journalistic sources [1]. Information extracted from Twitter has been leveraged for detection [2], extraction [3] and analysis of real-life events [4]. Humongous volume of tweets is posted every day with a considerable proportion of it being related to real-life events. This prolific event-specific content in Twitter makes it a promising ground for performing event analytics. Platforms like seen.co[[1]](#footnote-1), TwitterStand[[2]](#footnote-2), twitris[[3]](#footnote-3), Truthy[[4]](#footnote-4), and TweetTracker[[5]](#footnote-5) have developed techniques for providing analytics and organization of information related to different local and global real-life events.

Majority of the event related content shared in Twitter are personal updates of the users, pointless babbles, and spams. On the other hand, there are tweets that present high quality informative and newsworthy content. Table 1 presents examples of different types of tweets shared during events. Timely identification and tracking of high quality and informative tweets aids in gaining valuable information related to the events. It also enhances the experience of users interested in exploring the vital information related to the event and draw actionable insights. Journalists can gain by discovering high quality newsworthy tweets and sources producing them. Event managers and marketers can get benefited by identification of insightful information shared by the users as well as track highly interested users sharing informative content. Such a facility can also aid in cybersecurity, increase situational awareness and enable governments to take strategic actions related to events. Motivated by different use cases as stated above, in this paper we present the design and implementation of a framework that enables collection, extraction and management of high quality identity information related to real-life events from Twitter.

Table 1. Examples of different types of event related tweets.

|  |
| --- |
| *Thanks for the memories Sochi! I’ve had the time of my life #Sochi2014 #sochiselfie* [*http://t.co/DqkLEaAMpo*](http://t.co/DqkLEaAMpo)(**personal update) – low quality and non-informative, Event: Sochi Olympics 2014** |
| *Ted Cruz is a dangerous man. Crazy and gaining support. Megalomaniac leaders are bad, mkay. #CPAC #politics #joke)* (**pointless babble**) – **low quality and non-informative, Event: CPAC 2014** |
| *New post: Sochi Was For Suckers - Laugh Studios/ http://t.co/cWQJCBp3Ow #lol #funny #rofl #funnypic #fail #wtf.* (**spam**) – **low quality and non-informative, Event: CPAC 2014** |
| *In #Sochi, the Dutch are dominating the overall Olympic medal count* [*http://t.co/jMR1WUqEK4*](http://t.co/jMR1WUqEK4) *(Reuters)* [*http://t.co/dAfDhEgTGA*](http://t.co/dAfDhEgTGA). (**newsworthy and event-specific**) – **high quality and informative, Event: Sochi Olympics 2014** |
| *“@Breaking911: BREAKING NOW: #NYPD OFFICER INJURED ON THE BROOKLYN BRIDGE BY PROTESTERS THROWING ITEMS AT OFFICERS #MillionsMarchNYC” Great.* (**newsworthy and event-specific**) – **high quality and informative, Event: Millions March NYC 2015** |

Identity information of an event can be defined as a set of attribute values for that event along with a set of distinct rules that allow that event to be distinguished from all other events. The definition is inspired from the domain of entity resolution [5], where entity identity is one of the important factors for distinguishing one entity from another in a given context. In this work we present the implementation of a framework by considering events as entities. We take Sydney Siege crisis of 2015 as an instance of a real-life event. Grounding our work on the basic tenets of information quality we make the following primary contributions in this paper:

* Explain the implementation of a framework for collecting, extracting and managing event identity information, enabling tracking of updated event information and providing valuable event analytics from Twitter.
* Analyze information in 3.8 million event related tweets, and identify information units leading to high quality event information from short and colloquial textual content.
* Propose a novel Event Identity Information Structure (EIIS), capable of representing, updating and ranking identity information related to an event from Twitter.
* Present results obtained from tracking a real-life event (Sydney Siege Crisis, 2015).

**Related Work**

Entity resolution is one of the core topics that need attention in the literature behind the presented research. The other topics that are emphasized in the research are the tasks of identity management from social media, and identifying high quality information related to real-life events from Twitter.

Entity resolution has been known for five decades as the record linkage or the record matching problem in the statistics community [6]. The term Entity Resolution (ER) was first proposed in the research published by the Stanford InfoLab along with a generic model known as the Stanford Entity Resolution Framework (SERF) [7]. Historically, the focus of ER has been in developing processes and algorithms for determining if two references to an entity are equivalent. The Felligi Sunter model, SERF and the Talburt Wang model [8] are some of the prominent ones for conducting ER process. All the above models were based on matching algorithms working either at the record level or attribute level, and were tested with structured data in the relational databases. With the rise of big data, the modern trend is to perform entity resolution process in large volumes of unstructured data and scale it horizontally [9]. Although entity resolution from social media does involve processing of unstructured data, yet it has certain nuances. Most of the work in this regard so far has been in the area of entity extraction. Our effort of developing an *Event Identity Information Management* framework from short, informal and low quality textual content related to events, produced in Twitter would be a pioneering effort in the field of entity resolution and reference tracking, and would create new avenues of research.

Traditionally, identity resolution has been a subject of system administration and management of user identities in large organizations. For the first time [10] 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 [11, 12, 13]. 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 [14]. To our knowledge, we are the first to propose a framework for collecting and extracting identity information of real-life events from Twitter’s unstructured and informal text, and use the concepts of entity identity management and entity resolution for persistently managing their identities with respect to time.

Different web based platforms rank information shared in Twitter[[6]](#footnote-6). Recency of tweets, popularity based on retweet counts, authority of users and content relevance are the dominant factors used for ranking, in these platforms. A study of different state-of-the-art features commonly used for ranking tweets has been documented by [15]. Seen[[7]](#footnote-7) is a new state-of-the-art platform that uses a proprietary algorithm named *SeenRank* for ranking tweets in terms of event-specific information content for presenting event highlights and summaries to its users. In this work, we consider *SeenRank* as one of our baselines and compare our ranking with it. Apart from the existing real-world search applications, several adaptations of *PageRank* [16] have been proposed by the scientific community for ranking tweets and users in Twitter [17, 18, 19]. Various learning to rank approaches have been used for ordering tweets retrieved for a given query in terms of their relevance and quality [20, 21, 22]. Our work is different from all such approaches. One of the components of our presented framework simultaneously ranks different event-specific information units extracted from tweets and propagates their ranked scores to the tweets related to an event for identifying high quality informative tweets. The ranked information units are used as identity information for the event and helps in further tracking of high quality event related content from Twitter and to provide event analytics.

**Twitter**

Twitter is a social media website that allows its users to post short messages with a limitation of 140 characters. Along with posting plain textual content, the users also share images, videos and external URLs related to the content of the tweet. Twitter also allows its users to share content produced by other users, and such an activity is known as *retweeting*. The retweeted posts are preceded by special characters, ‘RT’. The messages are normally produced by a single user and are consumed by many. The consumers of information produced by an individual user in the context of Twitter are known as *followers*, and the user whom the other users follow is considered as their *friend*. A user produces content that might be of interest to him or his followers. On the other hand, a user generally follows another user who produces content of his interest.

The users in Twitter also participate in conversations among themselves and in loosely bound communities built around a topic. Whenever a user refers to another user in such conversations, they generally use the ‘@’ symbol followed by the username, commonly known as *user mentions*. The messages are contextualized for a topic by using a hashtag. Hashtags are a sequence of characters in any language prefixed by the symbol ‘#’ (e.g. #iciq2015). They are widely used by the users to join conversations related to a topic and to make tweets related to a topic easily searchable by other users. They act as strong identifiers of a topic in Twitter [23]. Any event related tweet can be considered to be a tweet related to a topic, which is the event itself. Therefore, while tweeting about real-life events, the users also use hashtags in order to post event-specific content.

Large amount of event related content generated in Twitter at great velocity creates the problem of information overload for the users interested in consuming the content. Due to brevity, noisiness, idiosyncratic language, unusual structure and ambiguous representation of discourse in Twitter it becomes extremely challenging for automated information mining tools and techniques to extract high quality information from tweets [24]. Information extraction tasks using state-of-the-art natural language processing techniques, often give poor results for tweets [25]. Abundance of link farms, unwanted promotional posts and nepotistic relationships between content creates additional challenges for our framework [26]. We devise novel techniques for overcoming these challenges and mine event identity information from tweets for identifying and analyzing high quality informative content related to events.

**Problem Definition**

Events have been defined from various perspectives and in different contexts. In the context of our work we adopt a definition similar to [27]. An event is defined as a real-world occurrence  with an associated time period, and a time ordered stream of tweets , of substantial volume, discussing the occurrence of the event and posted in time . Given such a scenario we present a formal definition of the problem that our proposed framework solves in this paper:

**Problem:** *Given a pre-specified finite set of real-life events* **** *with a pre-specified event (Ei* ∈ *E)* *generating a finite set of references* *****in social media, and a finite set of Event Identity Information Structures* **** *corresponding to each real-life event (Ei* ∈ *E), the problem is to resolve references of an event (Ei), and to persistently extract, store and manage identity information of the event in its corresponding Event Identity Information Structure  for identification of high quality insightful information related to the event over time period* *.*

**Event Identity Information Management**

Towards solving the specified problem we propose an *Event Identity Information Management* (EIIM) Framework comprising of different components and processes as shown in Figure 1. The various components of the framework go through cycles of interactions with each other over time, which is known as ***EIIM life cycle***. At the heart of the framework lies the *Event Identity Information Structure* (**EIIS**), which manages the identity information related to a particular event. Next, we present a detailed description of each component and processes of the EIIM life cycle. Whenever necessary we consider the tweets collected for a real-life event Sydney Siege Crisis 2015, in order to evaluate and demonstrate the results obtained from different components of the framework.

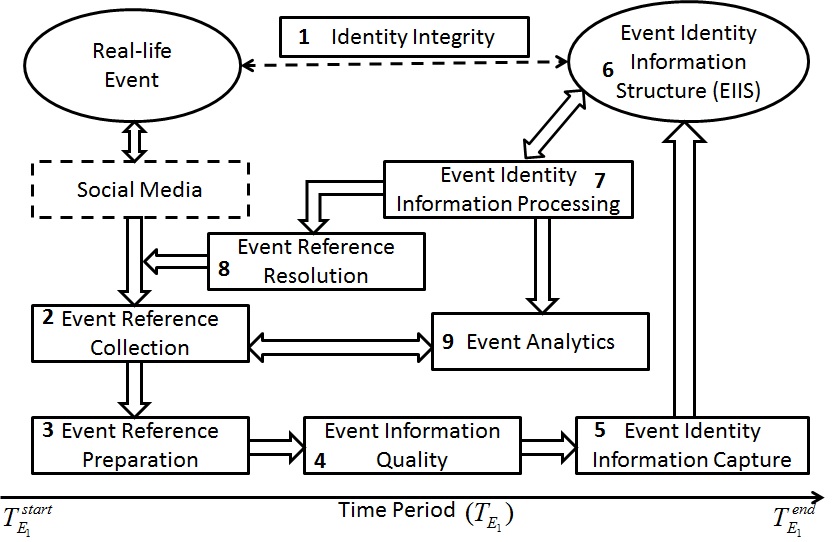


Figure 1. Event Identity Information Management Life Cycle

|  |  |
| --- | --- |
|  |  |

## 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[[8]](#footnote-8) 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.

## Event Reference Collection

This component allows the framework to collect event references from Twitter 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[[9]](#footnote-9)) is used for storage. The choice of MongoDb was also driven by its ability to scale horizontally and perform operations on large volumes of data.

For our case study we collected references of the Sydney Siege crisis event. Details of the collected event references are provided in Table 2. The tweets were collected over the given period of time, by providing a popular hashtag (#sydneysiege) to the Twitter streaming API[[10]](#footnote-10).

|  |  |
| --- | --- |
| **Event** | Sydney Siege Crisis |
| **Hashtag** | #sydneysiege |
| **No. of Tweets** | 398204 |
| **Time Period** | 15th Dec, 2014, 07:21:16 UTC  to  15th Dec, 2014; 22:46:45 UTC |

Table 2 Data Collected for Sydney Siege crisis.

The modules of this component work by interacting with the modules of the “*Identity Integrity*” component and “*Event Reference Resolution*”, for assigning a unique event identifier to the references of a particular event.

## Event Reference Preparation

Preprocessing the raw references is an important stage of any data intensive application. This component performs a series of data preparation steps on the collected event tweets in order to make them suitable for further processing by the other components of the EIIM life cycle. It performs deduplication of tweets using md5 hashing scheme. Redundant copies of a tweet are filtered out keeping a single copy in the database. Parts-of-speech tagging is done using the default POS tagger available in the NLTK[[11]](#footnote-11) module. A standard list of English stop words is used for eliminating the stop words from the tweet text. All the characters of a tweet are converted into lower case and special characters are removed. The tweets are tokenized into unigram tokens. User mentions, retweet symbol and URLs are removed during tokenization and are not considered as tokens.

A list of words expressing feelings in the internet, obtained from *wefeelfine.org* is used for detecting and extracting the feeling words from a tweet. Slang words commonly used in the internet and twitter specific slang publicly shared by FBI[[12]](#footnote-12) is combined together for compiling a list of English slang words. The modules use this list for detecting and extracting the slang words from the tweets, hashtags and text units. Retweet counts, favorite counts, verification information, user follower count, time information and expanded form of the URLs shared in the tweets are extracted from the metadata associated with each tweet, as retrieved using the Twitter API.

## Event Information Quality

This component examines the quality of information present in the tweets collected for the events. It segregates the references having high likelihood of containing good quality event related information from the ones that are less likely to contain or point to good quality information. In order to make a generic module for identifying high quality event related informative references we implemented a logistic regression classifier trained on a publicly available annotated dataset provided by [28]. The tweets labeled as ‘*related and informative*’ were assigned a score of 1 and all the other tweets labeled as ‘*related-but not informative*’, and ‘*not related*’ were assigned a score of 0. Table 3 lists the features extracted from each tweet. The choice of features was governed by previous works related to identifying high quality information from Twitter as already pointed in the *Related Work* section. 10-fold cross validation was performed resulting in a model with an accuracy of 76.64%. Table 4 lists the evaluation measures obtained while training the classifier.

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.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | No. of Unigram Tokens, No. of Stop Words, No. of Slang Words, No. of Feeling Words, No. of Hashtags, Has Url, No. of User Mentions, Length of the Tweet (No. of characters), No. of Unique Characters, No. of Special Characters, Favorite Count, Retweet Count, Formality [29], Is verified, No. of Nouns, No. of Adjectives, No. of Verbs, No. of Adverbs | | |  |  |  |  | | --- | --- | --- | --- | | **Logistic Regression**  **Model Performance** | **Precision** | **Recall** | **F-1 Score** | | *Non-informative* (0) | 0.70 | 0.49 | 0.57 | | *Informative* (1) | 0.78 | 0.90 | 0.84 | | Avg/Total  Accuracy = 76.64% | 0.76 | 0.77 | 0.75 | |
| **Table 3. Features extracted from a tweet for implementing the logistic regression model.** | **Table 4. 10-fold cross validation performance measures of the logistic regression model.** |

## Event Identity Information Capture

It is the component that aids in extracting *event identity information units* (explained later) from the already processed tweets and build the *Event Identity Information Structure* (EIIS) for an event. It also enables the framework to set a threshold between 0.0-1.0 for differentiating between high quality informative tweets from low quality non-informative ones related to an event. The *event identity information units* are then extracted from the high quality informative tweets.

In order to understand what might consist of the *event identity information units* that would represent the EIIS, we conducted a detailed analysis of 3.8 million tweets collected for three events. Details of the data collected are provided in Table 6. The data collection task was accomplished by *Event Reference Collection* component and was then preprocessed by the *Event Reference Preparation* component.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **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*** | | ***Non-informative*** | | |  | | --- | | 8.55 | | 3.55 | | |  | | --- | | 0.47 | | 0.77 | | |  | | --- | | 115.55 | | 69.92 | | |  | | --- | | 0.44 | | 1.23 | | |  | | --- | | 5.14 | | 1.78 | | |  | | --- | | 96.32% | | 1.04% | |
| **SXSW 2014** | |  | | --- | | ***Informative*** | | ***Non-informative*** | | |  | | --- | | 7.24 | | 3.08 | | |  | | --- | | 0.62 | | 0.91 | | |  | | --- | | 114.01 | | 62.64 | | |  | | --- | | 0.81 | | 0.94 | | |  | | --- | | 4.36 | | 1.52 | | |  | | --- | | 92.21% | | 0.34% | |
| **CPAC 2014** | |  | | --- | | ***Informative*** | | ***Non-informative*** | | |  | | --- | | 6.81 | | 3.55 | | |  | | --- | | 0.53 | | 0.9 | | |  | | --- | | 126.83 | | 88.65 | | |  | | --- | | 1.84 | | 2.04 | | |  | | --- | | 2.42 | | 2.04 | | |  | | --- | | 76.01% | | 0.68% | |

Table 5. Content characteristics of informative VS non-informative tweets related to events.

The logistic regression model developed for the *Event Information Quality* component was used for assigning scores to all the 3.8 million tweets in the dataset. The tweets getting a score greater than 0.7 were considered as instances of high quality informative tweets. Those getting a score lesser than 0.3 were considered as instances of low quality non-informative tweets. Average values of different content characteristics of the tweets were calculated. Top ten percent of the frequently occurring hashtags and nouns were considered as top hashtags and top nouns respectively, for the analysis. Some of the characteristics that were prominently different for informative and non-informative tweets are listed in Table 5.

As presented in the table, for all the three events, on an average the informative tweets are marked by a higher number of tokens per tweet and greater occurrence of top nouns. The average length of informative tweets is also more than the non-informative ones. The percentage of informative tweets having URLs is strikingly high. A greater use of slang words is observed in non-informative tweets. However, greater occurrence of top hashtags in non-informative tweets intrigued us to look into the content and obtain a detailed view of it. We observed that a lot of non-informative tweets have used popular hashtags with unrelated content and URLs directing to irrelevant information. This is typical of spam tweets as already reported by [30]. Although not shown due to space constraints, the average number of follower counts for users posting informative tweets was also observed to be higher than the ones posting non-informative ones.

The average number of feeling words used in informative tweets were also relatively higher than the feeling words used in the non-informative tweets.

Table 6. Details of data collected for analyzing event related data and identifying event identity information units.

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

The above observations gave us an idea of how high quality informative content related to events is produced in Twitter and the characteristics that differentiate them from low quality non- informative content. It is now intuitive that the informative tweets are more expressive, formal and lengthier, marked by higher presence of nouns. The high presence of nouns indicates that these tweets also contain information about people, places, organizations, etc, associated with the events, which is vital information about any event and is ideal for representing its identity. Due to the limitations imposed by Twitter on the number of characters in a tweet, the users tend to share URLs along with the textual content that might lead to more information about the event. Also, users with high follower counts tend to post informative tweets. This can also be concluded by the fact that as they have more followers they are encouraged to share informative content. Conversely, since they share informative content they are followed by a large number of other users interested in the content shared by them.

Based on the above analysis we decided to build the EIIS for an event composed of the following *event identity information units*:

* A finite set of hashtags ****used for annotating the tweets related to**.**
* A finite set of text units ****used for expressing textual content related to**.**
* A finite set of users ****tweeting about the event**.**
* A finite set of URLs **** shared in the tweets related to**.**
* A finite set of tweets **** related to the event**.**

Our decision to include the hashtags was based on the fact that the hashtags act as pseudo identifiers of event related information as already pointed out earlier. For our current implementation we use nouns as the text units. We also filter out the nouns and hashtags containing slang words in order to extract only good quality information. Different modules in this component are responsible for processing the tweets having a certain likelihood of containing high quality information as obtained from the previous component and extracting the above mentioned sets of *event identity information units* from them. The extracted *event identity information units* are persistently stored and metadata related to them are maintained as well as updated from time to time that is configurable.

## Event Identity Information Structure

This is the component that maintains a persistent EIIS for each individual event tracked by the framework and updates the metadata of the EIIS after a fixed interval of time which is configurable. We propose a novel graph structure for representing the event identity information units extracted by the previous component and call it *EventIdentityInfoGraph* as shown in Figure 2. The graph is dynamically generated from the extracted units of the previous component that acts as the vertices of the graph, after a fixed interval of time representing the snapshot of the *event identity information units* for a particular time period. Each vertex of the graph is assigned an initial event-specific score reflecting the importance of the vertices for a specific event. The score assigned to the tweets in the *Event Information* Quality component are used for the tweets. Due to the limited scope and focus of the paper, details of the calculations for the other type of vertices are not explained. The relationships between the heterogeneous vertices of the graph have special semantics that enables the framework to rank them in the next component. The semantics represent the following intuitive mutually reinforcing relationships between the vertices:

For an event ****

* A tweet is considered to be highly informative if it is strongly associated with (a) highly informative hashtags, (b) highly informative text-units, (c) highly informative users, (d) highly informative URLs.
* A hashtag is considered to be highly informative if it is strongly associated with (a) highly informative tweets, (b) highly informative text-units, (c) highly informative users, (d) highly informative URLs.
* A text unit is considered to be highly informative if it is strongly associated with (a) highly informative hashtags, (b) highly informative tweets, (c) highly informative users, (d) highly informative URLs.
* A user is considered to be highly informative if it is strongly associated with (a) highly informative hashtags, (b) highly informative tweets, (c) highly informative text units, (d) highly informative URLs.
* A URL is considered to be highly informative if it is strongly associated with (a) highly informative hashtags, (b) highly informative tweets, (c) highly informative users, (d) highly informative text units.

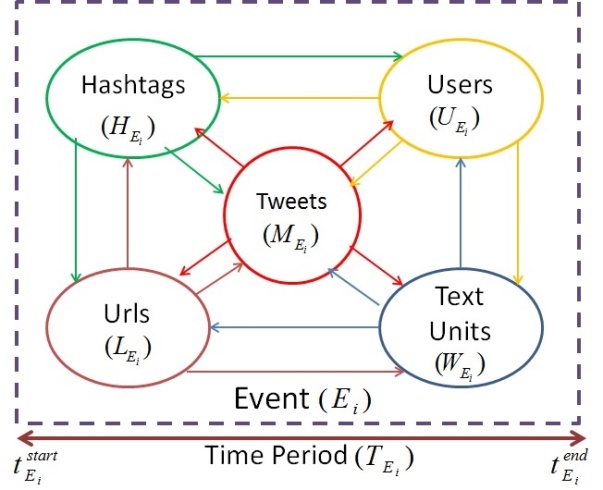


Figure 2 EventIdentityInfoGraph

We also assign a score to the edges connecting the vertices that quantify the semantics of their relationships. Due to the reasons already stated above, we don’t show the detailed calculations.

## Event Identity Information Processing

This component leverages the mutually reinforcing relationships represented by the *EventIdentityInfoGraph* of the previous component and ranks the vertices of the graph using a novel algorithm *EventIdentityInfoRank*, implemented by us. The algorithm is an iterative graph ranking algorithm that propagates the scores of each vertex to the adjoining vertices based on the weights assigned to the edges of the graph. After satisfying convergence criteria the algorithm converges giving the following sets as its output:

* A finite set of hashtags ****ranked in terms of event-specific informativeness.
* A finite set of text units **** ranked in terms of event-specific informativeness.
* A finite set of users ****ranked in terms of event-specific informativeness.
* A finite set of URLs ****ranked in terms of event-specific informativeness.
* A finite set of tweets ****ranked in terms of event-specific informativeness.

The ranking also assigns a final score to each of the vertices of the graph. We don’t provide the details of the algorithm as it is beyond the scope of the paper. But we provide the performance of the algorithm in identifying high quality informative tweets related to the Sydney Siege crisis event.

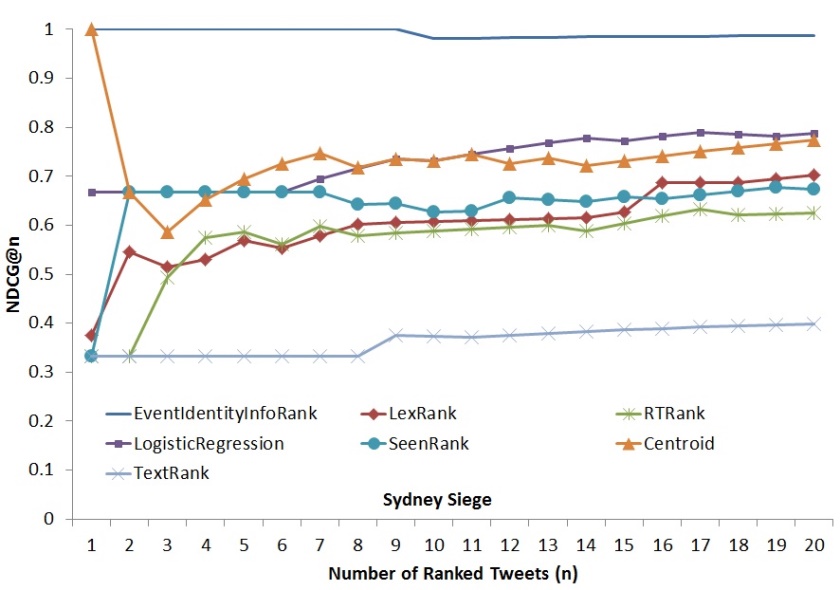


Figure 3. NDCG performance of different approaches in ranking tweets in terms of event-specific information for the Sydney Siege crisis event

We evaluated our algorithm against 6 other baseline techniques, as shown in Figure 3. The baseline techniques were: (a) LexRank [31], (b) TextRank [32], (c) RTRank, which is based on the number of retweets a tweet has, (d) Centroid [33], (e) SeenRank[[13]](#footnote-13) and (f) Logistic Regression Model used in the *Event Information Quality* component. We used our own logistic regression model as one of the baselines as we wanted to make sure that the event-specific ranking of tweets are different from a generic ranking of tweets only based on information quality. We also wanted to see how well *EventIdentityInfoRank* performs better than the initial information quality scores assigned to the tweets.

A subset of tweets for the event for a given time period (one hour) was selected. The choice of the time period was made on the basis of the intersection of the time period of the tweets collected by us and that provided by Seen for the same event. There were 37429 tweets. We obtained the ranked tweets for all the seven approaches. These ranked results were then annotated on an informativeness-scale of 1 to 3 (1 being least informative and 3 the most informative) by three graduate students as independent annotators. Necessary background of the event was given to the annotators. The annotators browsed the first hundred ranked tweets for all the seven approaches and assigned each of those tweets a rank from among the three ranks 1, 2 and 3. Thereafter, we computed Inter Indexer Consistency (IIC) values for the annotations of the datasets. The IIC value obtained was 0.83. The IIC value falls in the acceptable range of accuracy of annotations.

After being assured about consistency and accuracy of annotations, we moved to compute the Normalized Discounted Cumulative Gain (NDCG) [34] values at each of the hundred recall levels. This has been done for all the seven approaches. Figure 3 shows the NDCG curves up to 20 recall levels for all the seven approaches. It is quite evident from the figure that our approach outperforms all the baselines including the state-of-the-art approach of *SeenRank*. The algorithm is also robust against spam content and can simultaneously rank hashtags, text units, users, URLs and tweets related to an event in terms of their informativeness. The algorithm is also designed to be implemented in a distributed environment. Further progress is being currently made in order to deploy the algorithm in a Hadoop setup using the MapReduce paradigm.

## 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 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.

## Event Analytics

The outputs of the *Event Identity Information Processing* component after processing the *EventIdentityInfoGraph* of a particular event is used for generating different analytics related to the event for the chosen time period, by this component. Some of the immediately available analytics are shown in Table 6 and Table 7. The framework is capable of generating a variety of event analytics and can be extended to accommodate more options.

Table 6. Information gleaned from the Event Identity Information Processing component

|  |  |
| --- | --- |
| **Top Ten Event-specific Informative Text Units** | police, sydney, reporter, lindt, isis, nsw,  commissioner, australia, catherine,  martin |
| **Top Ten Event-specific Informative Hashtags** | #sydneysiege, #SydneySiege, #Sydneysiege, #MartinPlace, #9News,  #SydneyHostageCrisis, #Sydney, #Lindt, #ISIS, #SYDNEYSIEGE |
| **Top Five Informative Urls** | 1. <http://www.bbc.co.uk/news/world-australia-30474089>, 2. <http://www.cnn.com/2014/12/15/world/asia/australia-sydney-hostage-situation/index.html>, 3. <http://rt.com/news/214399-sydney-hostages-islamists-updates>, 4. <http://edition.cnn.com/2014/12/15/world/asia/australia-sydney-siege-scene/index.html>, 5. <https://vine.co/v/O6lgplBD1Uu>, |
| **Excerpts from Top Five Event-specific Informative Tweets** | 1. RT @faithcnn: Hostage taker in Sydney cafe has demanded 2 things: ISIS flag &amp; phone call with Australia PM Tony Abbott #SydneySiege http://… 2. Aussie grand mufti &amp; 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 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 |

Table 7. Tweets from top three informative users posting about Sydney Siege crisis.

|  |  |
| --- | --- |
| **Top Five Event-specific Informative Users** | **Three Randomly Selected Tweet Excerpts** |
| **User 1**  **Total no. of event related tweets by the user:** 41 | 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. RT @DR\_SHAHID: Hostage taker demands delivery of an #ISIS flag and a conversation with Prime Minister Tony Abbott <http://t.co/xTSDMKCPcD> 3. RT @SkyNewsBreak: Update - New South Wales police commissioner confirms five hostages have escaped from the Lindt cafe in Sydney #sydneysi… |
| **User 2**  **Total no. of event related tweets by the user:** 33 | 1. RT @smh: NSW Police Deputy Commissioner Catherine Burn will hold a press conference to update on the #SydneySiege at 6.30pm. 2. RT @Y7News: Helpful travel advice for commuters heading out of #Sydney’s CBD this evening - http://t.co/aQx2lvSosm #sydneysiege 3. RT @hughwhitfeld: British PM David Cameron informed of #sydneysiege .. UK Foreign Office is in touch with Aus authorities |
| **User 3**  **Total no. of event related tweets by the user:** 32 | 1. RT @RT\_com: #SYDNEY: Gunman tall man in late 40s, dressed in black – eyewitness http://t.co/m51P8dUPhB #SydneySiege <http://t.co/NvJzFsGrFN> 2. RT @NewsAustralia: 2GB's Ray Hadley claims hostage takers in #SydneySiege "wants to speak to Prime Minister Abbott live on radio." 3. RT @BBCWorld: "Profoundly shocking" -Australia PM Tony Abbott delivers second #sydneysiege statement. MORE: http://t.co/VaKt3ZpRZR http://… |

The entire EIIM life cycle as explained above is a controlled process for each individual event monitored by the framework. It is bootstrapped by providing a suitable hashtag for respective events that enables collection of tweets related to that event from Twitter. From there onwards, an event goes through the EIIM life cycle and passes through its different components (1-9) as shown in Figure 1, until it is terminated by human intervention.

# Conclusion and Future Work

We conclude our work by providing some final comments for our framework that shows its efficacy and usefulness. It also shows its readiness for implementing a fully functional system for a well-defined use case. Our thoughts and comments for different criterias are documented in Table 8.

Table 3 Concluding Comments on different evaluation criterias

|  |  |
| --- | --- |
| **Criteria** | **Comments** |
| Functional Completeness | The framework contains all the components as required to solve the problem and is functionally complete. |
| Functional Validity | All the components of the framework are correct and perform the functions as expected from them. |
| Usability | The framework can be easily extended and used for a variety of purposes in the domain of real-life events in different social media channels. |
| Test Coverage | All the components of the framework are tested and necessary data when available has been provided during explanation of each component. |
| Practical Utility | The framework has immense practical utility and can be used for *event analysis and monitoring, social media data integration, marketing campaigns, product launch monitoring from social media, search engine for event related information, journalistic applications from social media, Customer Relationship Management from social media, event based recommendations from social media,* etc. |
| Performance efficiency | The efficiency of the framework against state-of-the-art techniques has been demonstrated in the paper. |
| Effectiveness | The effectiveness of the framework against state-of-the-art techniques has been demonstrated in the paper. The raw analytics in Table 6 and Table 7 also show the effectiveness of the components. |
| Efficiency | The efficiency of the framework is currently under test. It is being improved on a regular basis for making it suitable for big data processing. Implementation of the framework using big data processing tools is underway. When completed it would make the framework extremely efficient in terms of time space complexity. |

In this paper we introduced the idea of Event Identity Information Management (EIIM) from Twitter and discussed how it could be used for tracking unstructured references related to a particular event. We also introduced the Event Identity Information Management life cycle and explained its various components. We pointed to our novel contributions in each of the component and showed the effectiveness and performance of our techniques against the state-of-the-art baselines.

As a future work we plan to scale the presented framework and make it capable of processing huge amount of data generated in real-time. We are also interested to expand our horizons and delve into other social media channels. It would certainly be possible to integrate the information related to an event from other social media channels as long as the references have the basic *event identity information units* as needed by the EIIM framework. We also plan to integrate more capabilities in the event analytics component, which includes real-time analytics, event summarization and recommendations.

## REFERENCES

1. Phelan, Owen, Kevin McCarthy, and Barry Smyth. "Using twitter to recommend real-time topical news." Proceedings of the third ACM conference on Recommender systems. ACM, 2009.
2. Sakaki, Takeshi, 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 (2013): 919-931.
3. Popescu, Ana-Maria, Marco Pennacchiotti, and Deepa Paranjpe. "Extracting events and event descriptions from Twitter." Proceedings of the 20th international conference companion on World wide web. ACM, 2011.
4. Purohit, Hemant, and Amit P. Sheth. "Twitris v3: From Citizen Sensing to Analysis, Coordination and Action." ICWSM. 2013.
5. Lim, E. P., et al. (1993, April). Entity identification in database integration. In Data Engineering, 1993. Proceedings. Ninth International Conference on (pp. 294-301). IEEE.
6. Winkler, William E. "The state of record linkage and current research problems." Statistical Research Division, US Census Bureau. 1999.
7. Benjelloun, Omar, et al. "Generic entity resolution in the serf project." IEEE Data Engineering Bulletin, June 2006 Issue (2006).
8. John R Talburt. Entity resolution and information quality. Elsevier, 2011. Arkady Maydanchik. Data quality assessment. Technics publications, 2007.
9. Lars, Kolb, et al. "Dedoop: efficient deduplication with Hadoop." Proceedings of the VLDB Endowment 5.12 (2012): 1878-1881.
10. Zhou, Yinle, et al. Entity identity information management (eiim). In International Conference on Information Quality (ICIQ-11), Adelaide, Australia, pages 327{341, 2011.
11. Bouquet, Paolo, et al. "Entity-centric Social Profile Integration." Proceedings of the International Workshop on Linking of User Profiles and Applications in the Social Semantic Web (LUPAS 2010). 2010.
12. Bortoli, Stefano, et al. "Foaf-O-Matic-Solving the Identity Problem in the FOAF Network." SWAP. 2007.
13. Raad, Elie, et al. "User profile matching in social networks." Network-Based Information Systems (NBiS), 2010 13th International Conference on. IEEE, 2010.
14. Bouquet, Paolo, et al. "OkkaM: Towards a Solution to the``Identity Crisis''on the Semantic Web." SWAP. Vol. 201. 2006.
15. Damak, Firas, et al. "Effectiveness of State-of-the-art Features for Microblog Search." Proceedings of the 28th Annual ACM Symposium on Applied Computing. ACM, 2013.
16. Page, Lawrence, et al. "The PageRank citation ranking: Bringing order to the web." (1999).
17. Weng, Jianshu, et al. "Twitterrank: finding topic-sensitive influential twitterers." Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010.
18. Tunkelang, D. "A twitter analog to pagerank." The Noisy Channel (2009).
19. Hallberg, V., et al. "An adaptation of the PageRank algorithm to Twitter world." (2012).
20. Duan, Yajuan, et al. "An empirical study on learning to rank of tweets."Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010.
21. McCreadie, Richard, and Craig Macdonald. "Relevance in microblogs: Enhancing tweet retrieval using hyperlinked documents." Proceedings of the 10th conference on open research areas in information retrieval. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2013.
22. Vosecky, Jan, Kenneth Wai-Ting Leung, and Wilfred Ng. "Searching for quality microblog posts: Filtering and ranking based on content analysis and implicit links." Database Systems for Advanced Applications. Springer Berlin Heidelberg, 2012.
23. Laniado, David, and Peter Mika. "Making sense of twitter." The Semantic Web–ISWC 2010. Springer Berlin Heidelberg, 2010. 470-485.
24. Bontcheva, Kalina, et al. "TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text." RANLP. 2013.
25. Ritter, Alan, Sam Clark, and Oren Etzioni. "Named entity recognition in tweets: an experimental study." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.
26. Gayo-Avello, Daniel. "Nepotistic relationships in twitter and their impact on rank prestige algorithms." Information Processing & Management 49.6 (2013): 1250-1280.
27. Becker, Hila, Mor Naaman, and Luis Gravano. "Beyond Trending Topics: Real-World Event Identification on Twitter." ICWSM 11 (2011): 438-441.
28. Olteanu, Alexandra, et al. "CrisisLex: A lexicon for collecting and filtering microblogged communications in crises." In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM" 14). No. EPFL-CONF-203561. 2014.
29. Alejandro, Mosquera, and Moreda Paloma. "The use of metrics for measuring informality levels in web 2.0 texts." (2011).
30. Yardi, Sarita, and Danah Boyd. "Dynamic debates: An analysis of group polarization over time on twitter." Bulletin of Science, Technology & Society30.5 (2010): 316-327.
31. Erkan, Günes, and Dragomir R. Radev. "LexRank: graph-based lexical centrality as salience in text summarization." Journal of Artificial Intelligence Research (2004): 457-479.
32. Mihalcea, Rada, and Paul Tarau. "TextRank: Bringing order into texts." Association for Computational Linguistics, 2004.
33. Becker, Hila, Mor Naaman, and Luis Gravano. "Selecting Quality Twitter Content for Events." ICWSM 11 (2011).
34. Järvelin, Kalervo, and Jaana Kekäläinen. "Cumulated gain-based evaluation of IR techniques." ACM Transactions on Information Systems (TOIS) 20.4 (2002): 422-446.

1. http://seen.co [↑](#footnote-ref-1)
2. http://twitterstand.umiacs.umd.edu/ [↑](#footnote-ref-2)
3. http://twitris.knoesis.org/ [↑](#footnote-ref-3)
4. http://truthy.indiana.edu/ [↑](#footnote-ref-4)
5. http://tweettracker.fulton.asu.edu/ [↑](#footnote-ref-5)
6. http://mashable.com/2009/04/22/twitter-search-services [↑](#footnote-ref-6)
7. http://seen.co [↑](#footnote-ref-7)
8. http://docs.mongodb.org/manual/reference/object-id/ [↑](#footnote-ref-8)
9. http://www.mongodb.org/ [↑](#footnote-ref-9)
10. https://dev.twitter.com/streaming/overview [↑](#footnote-ref-10)
11. http://nltk.org [↑](#footnote-ref-11)
12. https://www.documentcloud.org/documents/1199460-responsive-documents.html#document/p1 [↑](#footnote-ref-12)
13. Tweets in Seen.co is ranked according to their proprietary algorithm SeenRank and the scores are available in the response of their API found at (http://developer.seen.co/) We used a python wrapper freely available at <https://github.com/dxmahata/pySeen> for collecting data from Seen.co [↑](#footnote-ref-13)