From Chirps to Whistles Discovering Event-specific Informative Content from Twitter

Debanjan Mahata
Department of Information Science
University of Arkansas at Little Rock
Little Rock, USA
dxmahata@ualr.edu

John R. Talburt
Department of Information Science
University of Arkansas at Little Rock
Little Rock, USA
jrtalburt@ualr.edu

Vivek Kumar Singh
Department of Computer Science
South Asian University
New Delhi, India
vivek@cs.sau.ac.in

ABSTRACT

Twitter has brought a paradigm shift in the way we produce and curate information about real-life events. Huge volumes of user-generated tweets are produced in Twitter, related to events. Not, all of them are useful and informative. A sizable amount of tweets are spams and colloquial personal status updates, which does not provide any useful information about an event. Thus, it is necessary to identify, rank and segregate event-specific informative content from the tweet streams. In this paper, we develop a novel generic framework based on the principle of mutual reinforcement, for identifying event-specific informative content from Twitter. Mutually reinforcing relationships between tweets, hashtags, text units, URLs and users are defined and represented using *TwitterEventInfoGraph*. An algorithm - *TwitterEventInfoRank* is proposed, that simultaneously ranks tweets, hashtags, text units, URLs and users producing them in terms of event-specific informativeness by leveraging the semantics of relationships between each of them as represented by *TwitterEventInfoGraph*. Experiments and observations are reported on four million (approx) tweets collected for five real-life events, and evaluated against popular baseline techniques showing significant improvement in performance.

Keywords

twitter, mutual reinforcement, event, information retrieval, ranking, event-specific information

1. INTRODUCTION

Social media platforms provide multiple venues to people for sharing first-hand experiences and exchange information about real-life events. Twitter is one such platform that has become an indispensable source for disseminating news and real-time information about current events. It is a microblogging application that allows its users to post short messages of 140 characters known as tweets, from a variety of internet enabled devices. Studies have shown the importance of Twitter as a news circulation service [22], and a source for gauging public interest and opinions [19]. It's efficacy as a real-time citizen-journalistic source of information has been recently harnessed in detection, extraction and analysis of real-life events [27, 23, 24].

Users not only post plain textual content in their messages but also share URLs, linking to other external websites, images and videos. Apart from creating new content, the users also share content produced by others. This activity is known as *retweeting*, and such tweets are preceded by special characters '*RT*'. The messages are normally written by a single person and are read by many. The readers in this context are known as *followers*, and the user whom they follow is considered as their *friend*. Any user with good intent either share messages that might be of interest to his followers, or for joining conversations on topics of his interest. The '@' symbol followed by the username commonly known as *user mentions*, is used for mentioning other users in tweets for initiating conversations.

The concise and informal content of a tweet is often contextualized by the use of a crowdsourced annotation scheme called *hashtags*. Hashtags are a sequence of characters in any language prefixed by the symbol '#' (for e.g. #websci2015). They are widely used by the users for categorizing the content based on a topic, join conversations related to a topic, and to make the tweets easily searchable by other interested users. They also act as strong identifiers of topics [13]. When tweeting about real-life events the users also tend to use hashtags in order to post event-specific content. For e.g. '#Egypt' and '#Jan25', were among the most popular hashtags in Twitter used for spreading, organizing and analyzing information related to 'Egyptian Revolution of 2011' [3].

284 million monthly users of Twitter posting 500 million tweets per day produces a variety of content[1]. A significant proportion of it are related to different real-life events (e.g, football matches, conferences, music shows, etc). Majority of this content are personal updates (e.g. *Thanks for the memories Sochi! I've had the time of my life #Sochi2014 #sochiselfie http://t.co/DqkLEaAMpo*), pointless babbles (e.g.

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

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 the context of this paper we call them as event-specific informative tweets.

Motivation: With the plethora of event related content being produced in Twitter, it becomes inconvenient for users to search and follow informative posts. This necessitates development of techniques that can identify and rank tweets in terms of their event-specific informativeness. In addition to the tweets, a backend automated system dedicated for processing, analyzing and presenting information from Twitter during an event, could get immensely benefitted from identification and ranking of event-specific informative hashtags, text units, users and URLs. This would enable the system to generate answers to questions like: *Who are the users producing large amount of event-specific informative content?. Which are the best hashtags and URLs to follow that would lead to high quality event-specific information?. Which are the best hashtags and text units to index for efficient retrieval of event-specific information?*. Such a system would further facilitate better consumption of content while exploring event information from Twitter. It could have a positive impact on triggering event-specific recommendations and efficient processing of information. It can act as a core component of event management, event summarization, event marketing and journalistic platforms leveraging Twitter.

Challenges: Apart from the problem of information overload, microblogging websites like Twitter pose challenges for automated information mining tools and techniques due to their brevity, noisiness, idiosyncratic language, unusual structure and ambiguous representation of discourse. 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. Due to the lack of explicit links between content shared in Twitter it is also difficult to implement and get useful results from ranking algorithms popularly used for web pages. Lastly, to our knowledge, there is an absence of techniques at present that is capable of simultaneously ranking and identifying event-specific informative tweets, hashtags, text units, users and URLs, with an ability to scale.

Objective and Contributions: The main objective of our work is to automatically identify and rank event-specific informative content posted in Twitter. Our primary hypothesis is that there are explicit cues available in the content of the tweets posted during an event for determining event-specific informativeness. Our approach is based on the *principle of mutual reinforcement* commonly used for summarization of textual documents. We build our methodology on the basic tenets of *Mutually Reinforcing Chains* [31], for ranking and identification of event-specific informative content in Twitter. We make the following contributions:

- analysis of informative and non-informative content in 3.8 million event related tweets;
- propose a generic framework based on principle of mutual reinforcement that takes into account the semantics of relationships between *tweets*, *hashtags*, *text units*, *URLs* and *users*, and represent them in a graph structure - *TwitterEventInfoGraph*;
- leverage the mutually reinforcing relationships in *TwitterEventInfoGraph* and develop a graph based iterative algorithm - *TwitterEventInfoRank*, for simultaneously ranking *tweets*, *hashtags*, *text units*, *users* and *URLs* in terms of event-specific informativeness;
- evaluate the algorithm against popular baselines and report its performance in identifying and ranking event-specific informative content from Twitter.

Next, we review work related to the presented paper.

2. RELATED WORK

Prior work relevant to ours is mainly related to, *ranking of tweets* and *summarization of Twitter content.* We present them in this section. Whenever appropriate, we also point to the literature behind the baseline techniques used for evaluating our work.

Ranking of Tweets: 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 [6, 17]. 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* [21] has been proposed by the scientific community for ranking tweets and users in Twitter [32, 29, 9]. Various learning to rank approaches have been used for ordering tweets retrieved for a given query in terms of their relevance and quality [7, 15, 30]. 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 [5]. They represented tweets of an event in a cluster and calculated the similarity of individual

[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

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.

Summarization of Twitter Content: 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 [11]. Summarization of tweets for sporting events was performed by [18] using the phrase graph algorithm [28]. The popularly used graph-based summarization algorithms are *LexRank* [8] and *TextRank* [16]. 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 (*TwitterEventInfoGraph*), which is novel and makes our work different from any prior work. 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.

Next, we present a detailed analysis of event related tweets that aids in gleaning through the nature of informative and non-informative content produced during real-life events. It also helps in understanding the difference between generic informative content and event-specific informative content, which is one of the factors that intrigues and motivates us to solve the problem as presented in this paper.

3. ANALYSIS OF EVENT RELATED TWEET CONTENT

Table 1: Details of data collected for analyzing event related tweet content.

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

Given the mechanisms of user interactions and content production in Twitter as explained in Section 1, we analysed 3.8 million (approx) English tweets produced during three real-life events. Details of the data related to the events, collected for conducting the analysis is presented in Table 1[5]. We provided a popular hashtag corresponding to each event to the Twitter streaming API[6] in order to collect the data over the indicated period of time. The text of the tweets were preprocessed and prepared for analysis (Refer Section 6.2 for details). One of the main intentions behind this analysis was to investigate the nature of content in informative and non-informative tweets and to understand if there is a difference between tweets rich in generic information and the ones with event-specific information.

Table 2: Tweet features for content informativeness.

Has Url, No. of words, No. of stopwords, No. of feeling words [7], No. of slang words, No. of hashtags, No. of user mentions, Tweet length (No. of characters), No. of unique characters, No. of special characters, Favorite count, Retweet count, Formality[8] [1], Is tweet verified, No. of nouns, No. of adjectives, No. of verbs, No. of adverbs, No. of pronouns, No. of interjections, No. of articles, No. of prepositions.

Table 3: Evaluation measures for logistic regression model.

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

Analysis of informative and non-informative content: Our first step was to segregate the tweets likely to have informative content from the non-informative ones. In order to do so we trained a logistic regression model on an annotated dataset [20], which is publicly available. 9729 English language annotated tweets were used for building the model. The tweets labeled as *related and informative* were assigned a score of 1 and all the other tweets labeled as *related - but not informative* and *not related* were assigned a score of 0. Table 2 lists the features selected for each tweet. The choice of the features was driven by previous studies pointed out in Section 2. 10-fold cross validation was performed resulting in a model with an accuracy of 76.64%. Refer Table 3 for evaluation measures.

The logistic regression model was then used for assigning informativeness score between 0 and 1 to all the tweets in the dataset, with 0 being least informative and 1 being most informative. Although, the model is developed on tweets related to disaster events, it has been shown by the authors [10] that the annotations could be generalized to any type of event. The tweets for each event were then separated into two subsets - *informative*, containing tweets scoring more than 0.7, and *non-informative*, containing tweets scoring less than 0.3. Average values of different content characteristics of tweets were calculated for both the subsets. The top 10% of the frequently occurring hashtags, nouns and URLs were considered as top hashtags, nouns and URLs for the analysis, respectively. Some of the characteristics that were noticeably different for informative and non-informative tweets are listed in Table 4.

For all the events, on an average, the informative tweets were marked by a higher number of words per tweet and

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

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

[7] http://wefeelfine.org

[8] Formality = (#nouns + #adjectives + #prepositions + #articles - #pronouns - #verbs - #adverbs - #interjections + 100)/2

Table 4: Content characteristics of informative and non-informative tweets related to events.

		Average content characteristics per tweet					
		# of words	# of slang words	Length	Top hashtags	Top nouns	Urls Percent
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	1.0	0.68%

greater occurrence of top nouns. The average length of informative tweets were also more than the non-informative ones. The percentage of informative tweets having URLs were strikingly high. As expected, a greater usage of slang words was observed in non-informative tweets. However, the greater occurrence of top hashtags in non-informative tweets urged us to look into the content. We found that a lot of non-informative tweets have used many popular hashtags with unrelated content and URLs pointing to irrelevant information. This is typical of spam tweets as already reported by [33]. Not shown due to space constraints, a larger average occurrence of feeling words and top URLs was observed in informative tweets. The average number of follower counts for users posting informative tweets was also observed to be higher than the ones posting non-informative ones.

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

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

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

Next, we present a formal definition of the problem that we solve in this paper.

4. PROBLEM STATEMENT

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

Events have been defined from various perspectives and in different contexts. In the context of our work we adopt a definition similar to [4]. **Event:** An event is defined as a real-world occurrence (E_i) with an associated time period T_{E_i} ($t^{start}_{E_i}$-$t^{end}_{E_i}$) and a time ordered stream of tweets M_{E_i}, of substantial volume, discussing about the event and posted in time T_{E_i}. The tweets are primarily composed of a set of hashtags (H_{E_i}) used for annotating the tweets ($\in M_{E_i}$), a set of text units (W_{E_i}) used for sharing textual information in the tweets ($\in M_{E_i}$), a set of URLs (L_{E_i}) linking to external sources related to the event and a set of users (U_{E_i}) posting the tweets ($\in M_{E_i}$).

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

- *tweets* $\hat{M}_{E_i}=\{m_1\geq...\geq m_i\geq m_j\geq...\geq m_n|i<j\}$,
- *hashtags* $\hat{H}_{E_i}=\{h_1\geq...\geq h_i\geq h_j\geq...\geq h_p|i<j\}$,
- *text units* $\hat{W}_{E_i}=\{w_1\geq...\geq w_i\geq w_j\geq...\geq w_r|i<j\}$,
- *URLs* $\hat{L}_{E_i}=\{l_1\geq...\geq l_i\geq l_j\geq...\geq l_t|i<j\}$,
- *users* $\hat{U}_{E_i}=\{u_1\geq...\geq u_i\geq u_j\geq...\geq u_s|i<j\}$,

ordered in decreasing order of its event-specific informativeness.

Next, we explain the methodology used for solving the problem.

5. METHODOLOGY

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

Table 5: Affinity scores of edges and event-specific initialization scores for nodes of TwitterEventInfoGraph

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

$P(h_i|w_j) = \frac{No.\ of\ tweets\ h_i\ and\ w_j\ occur\ together}{No.\ of\ tweets\ w_j\ occurs}$, $P(w_i|h_j) = \frac{No.\ of\ tweets\ w_i\ and\ h_j\ occur\ together}{No.\ of\ tweets\ h_j\ occurs}$, $P(h_i|l_j) = \frac{No.\ of\ tweets\ h_i\ and\ l_j\ occur\ together}{No.\ of\ tweets\ l_j\ occurs}$

$P(l_i|h_j) = \frac{No.\ of\ tweets\ l_i\ and\ h_j\ occur\ together}{No.\ of\ tweets\ h_j\ occurs}$, $P(h_i|u_j) = \frac{No.\ of\ tweets\ h_i\ and\ u_j\ occur\ together}{No.\ of\ tweets\ u_j\ occurs}$, $P(u_i|h_j) = \frac{No.\ of\ tweets\ u_i\ and\ h_j\ occur\ together}{No.\ of\ tweets\ h_j\ occurs}$,

$P(w_i|l_j) = \frac{No.\ of\ tweets\ w_i\ and\ l_j\ occur\ together}{No.\ of\ tweets\ l_j\ occurs}$, $P(l_i|w_j) = \frac{No.\ of\ tweets\ l_i\ and\ w_j\ occur\ together}{No.\ of\ tweets\ w_j\ occurs}$, $P(w_i|u_j) = \frac{No.\ of\ tweets\ w_i\ and\ u_j\ occur\ together}{No.\ of\ tweets\ u_j\ occurs}$,

$P(u_i|w_j) = \frac{No.\ of\ tweets\ u_i\ and\ w_j\ occur\ together}{No.\ of\ tweets\ w_j\ occurs}$, $P(u_i|l_j) = \frac{No.\ of\ tweets\ u_i\ and\ l_j\ occur\ together}{No.\ of\ tweets\ l_j\ occurs}$, $P(l_i|u_j) = \frac{No.\ of\ tweets\ l_i\ and\ u_j\ occur\ together}{No.\ of\ tweets\ u_j\ occurs}$,

$P(h_i|m_j) = P(m_i|h_j) = P(w_i|m_j) = P(m_i|w_j) = P(u_i|m_j) = P(m_i|u_j) = P(l_i|m_j) = P(m_i|l_j) = 1.0$, **Note:** $P(h_i \mid w_j)$ should be read as the probability of occurrence of hashtag h_i given the occurrence of the text unit w_j in the stream of tweets M_{E_i} related to event E_i collected over the time period T_{E_i}.

Event-specific initialization scores of vertices $\in H_{E_i}, W_{E_i}, U_{E_i}, L_{E_i}$:

$Score(h_i) = \frac{freq(h_i)}{max\{freq(h_1), freq(h_2), ..., freq(h_p)\}}$ (1) $Score(w_i) = \frac{freq(w_i)}{max\{freq(w_1), freq(w_2), ..., freq(w_r)\}}$ (2) $Score(u_i) = \frac{followers(u_i)}{max\{followers(u_1), ..., followers(u_r)\}}$ (3)

$Score(l_i) = \frac{freq(l_i)}{max\{freq(l_1), freq(l_2), ..., freq(l_r)\}}$ (4) where, $freq(h_i)$ is the frequency of occurrence of the i^{th} hashtag ($\in H_{E_i}$) in the stream of tweets M_{E_i}. Similarly, $freq(w_i)$ denotes the frequency of occurrence of the i^{th} text unit ($\in W_{E_i}$) and, $freq(l_i)$ denotes the frequency of occurrence of the i^{th} url ($\in L_{E_i}$). $followers(u_i)$ denotes the number of followers of user $u_i \in (U_{E_i})$.

5.1 TwitterEventInfoGraph

After the observations in the previous section we conclude that the informative tweets in general are characterized by wordiness, occurrences of URLs and are posted by users with high follower count. These characteristics are also the primary features that distinguish informative from non-informative content. Although, presence of hashtags is not a good indicator of informativeness, yet it is a strong identifier of a topic as already pointed by [13]. Popular hashtags for an event might be used maliciously. On the other hand, the presence of a popular hashtag in a wordy tweet consisting of words popular for the event, along with a popular URL, posted by an influential user is highly likely to contain event-specific content. Therefore, it is intuitive that given a stream of tweets for an event an optimal combination of event related popular text units (words, unigrams, bigrams etc), hashtags, and URLs, posted by an influential user in a tweet, is one of the key indicators for identifying event-specific informative content. It would be highly unlikely for a tweet to contain all of these and yet not convey useful event-specific information. Building on this intuition we model our framework based on the following assumptions:

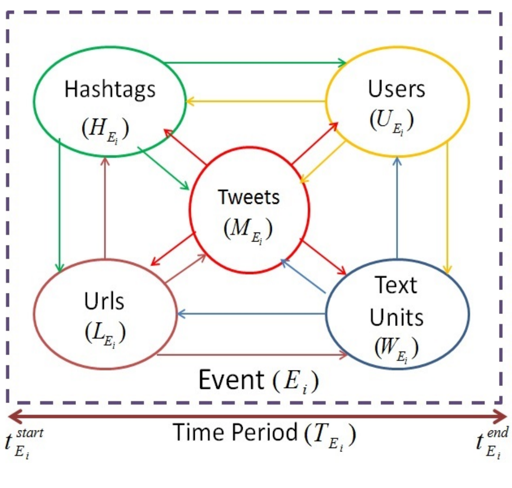

Figure 1: Mutual Reinforcement Chains in Twitter for an event.

For an event E_i

- a *tweet is an event-specific informative tweet* if it is strongly associated with: **(a)** *event-specific informative hashtags*, **(b)** *event-specific informative text units*, **(c)** *event-specific informative users*, **(d)** *event-specific informative URLs.*
- a *hashtag is an event-specific informative hashtag* if it is strongly associated with: **(a)** *event-specific informative tweets*, **(b)** *event-specific informative text units*, **(c)** *event-specific informative users*, **(d)** *event-specific informative URLs.*
- a *text unit is an event-specific informative text unit* if it is strongly associated with: **(a)** *event-specific informative tweets*, **(b)** *event-specific informative hashtags*, **(c)** *event-specific informative users*, **(d)** *event-specific informative URLs.*
- a *user is an event-specific informative user* if it is strongly associated with: **(a)** *event-specific informative tweets*, **(b)** *event-specific informative hashtags*, **(c)** *event-specific informative text units*, **(d)** *event-specific informative URLs.*
- a *URL is an event-specific informative URL* if it is strongly associated with: **(a)** *event-specific informative tweets*, **(b)** *event-specific informative hashtags*, **(c)** *event-specific informative text units*, **(d)** *event-specific informative users.*

The relationships for an event E_i as stated above, forms a *Mutual Reinforcement Chain* [31] for the event E_i as shown in Figure 1. We represent this relationship in a graph $\mathbf{G = (V,D)}$, which we call as *TwitterEventInfoGraph*, where $\mathbf{V} = \mathbf{M_{E_i}} \cup \mathbf{H_{E_i}} \cup \mathbf{W_{E_i}} \cup \mathbf{U_{E_i}} \cup \mathbf{L_{E_i}}$, is the set of vertices and $\mathbf{D}$ is the set of directed edges between different vertices.

Whenever two vertices are associated, there are two edges between them that are oppositely directed. Each directed edge is assigned a weight, which determines the degree of association of one vertex with the other. The weights for each edge is calculated according to the conditional probabilities given in Table 5.

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

Next, we explain *TwitterEventInfoRank*.

5.2 TwitterEventInfoRank

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

We first assign a event-specific score to all the vertices of the graph. Event-specific scores for vertices ($\in H_{E_i}, W_{E_i}, U_{E_i}, L_{E_i}$) are calculated using equations (1-4) as presented in Table 5. The tweets ($\in M_{E_i}$) are assigned an initial informativeness score as obtained from the logistic regression model explained in Section 3. The event-specific scores for vertices ($\in H_{E_i}, W_{E_i}, U_{E_i}, L_{E_i}$) and informativeness score for vertices ($\in M_{E_i}$) gives an initial ranking of all the vertices of *TwitterEventInfoGraph*. We aim to refine the initial scores and assign a final score for ranking the vertices by leveraging the mutually reinforcing relationships between them.

The relationships between two different subsets of vertices in graph $\mathbf{G}$ is denoted by an affinity matrix. For e.g., $\mathbf{A_{E_i}^{MH}}$ denotes the $\mathbf{M_{E_i}}-\mathbf{H_{E_i}}$ affinity matrix for event E_i, where $\mathbf{(i,j)^{th}}$ entry is the edge weight quantifying the association between i^{th} tweet ($\in M_{E_i}$) and j^{th} hashtag ($\in H_{E_i}$), calculated using Table 5. 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 $R_{E_i}^M$, $R_{E_i}^H$, $R_{E_i}^W$, $R_{E_i}^U$ and $R_{E_i}^L$ 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:

$$R_{E_i}^{M(k+1)}=A_{E_i}^{MM(k)}+A_{E_i}^{MH(k)}+A_{E_i}^{MW(k)}+A_{E_i}^{MU(k)}+A_{E_i}^{ML(k)} \quad (1)$$

$$R_{E_i}^{H(k+1)}=A_{E_i}^{HM(k)}+A_{E_i}^{HH(k)}+A_{E_i}^{HW(k)}+A_{E_i}^{HU(k)}+A_{E_i}^{HL(k)} \quad (2)$$

$$R_{E_i}^{W(k+1)}=A_{E_i}^{WM(k)}+A_{E_i}^{WH(k)}+A_{E_i}^{WW(k)}+A_{E_i}^{WU(k)}+A_{E_i}^{WL(k)} \quad (3)$$

$$R_{E_i}^{U(k+1)}=A_{E_i}^{UM(k)}+A_{E_i}^{UH(k)}+A_{E_i}^{UW(k)}+A_{E_i}^{UU(k)}+A_{E_i}^{UL(k)} \quad (4)$$

$$R_{E_i}^{L(k+1)}=A_{E_i}^{LM(k)}+A_{E_i}^{LH(k)}+A_{E_i}^{LW(k)}+A_{E_i}^{LU(k)}+A_{E_i}^{LL(k)} \quad (5)$$

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

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

Let

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

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

$$\Delta_{E_i}.R_{E_i}=\lambda.R_{E_i} \quad (6)$$

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

$$\overline{\Delta}_{E_i}=\alpha\hat{\Delta}_{E_i}+(1-\alpha)E \quad (7)$$

$$E=p\times[1]_{1\times k} \quad (8)$$

where $0\leq\alpha\leq1$ is set to 0.85 according to *PageRank*, and k is the order of $\hat{\Delta}_{E_i}$. We set $p=[1/k]_{k\times1}$ 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,

- 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.
- Then we assign

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

  and normalize $R_{E_i}^0$ such that $||R_{E_i}^0||_1=1$
- Apply power iteration method using the same parameters as used in PageRank with the convergence tolerance set at 1e-08 and $\lambda = 0.85$.
- We get the final rank vectors for each subset of the vertices ($R_{E_i}^M$,$R_{E_i}^H$,$R_{E_i}^W$,$R_{E_i}^U$,$R_{E_i}^L$) after convergence.
- 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.

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

The text units are generic units in the framework and can be changed according to specific requirements. Entities extracted from the textual content of tweets could be experimented, in place of nouns. Since the algorithm uses power iteration method for ranking the vertices of the graph, it could be easily made scalable using mapreduce paradigm [14]. We plan to work on it in the future and implement our framework using hadoop and mapreduce environment.

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

Next, we present details of experimental settings and evaluation.

6. EXPERIMENTAL SETTINGS AND EVALUATION

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

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

6.1 Data Collection

For implementing and evaluating our proposed algorithm we collected 455,131 tweets from two real-life events, 'Millions March NYC' and 'Sydney Siege', using Twitter Streaming API. Details of the dataset is presented in Table 6. Tweets for each event was collected over the given period of time, by providing a popular hashtag corresponding to each event to the Twitter streaming API. The events for the experiments are different from the events selected for initial analysis as the choice was driven by its availability in Seen.co event

Table 6: Details of data collected for the experiment.

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

database, whose ranking scores[9] are used as one of the baselines representing the state-of-the-art technique.

6.2 Data Preparation

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

We used the default parts-of-speech (POS) tagging module provided by NLTK library[10]. A standard list of english stop words was used for eliminating the stop words from tweet text. All the characters of the tweets were converted to lower case. The tweets were tokenized after detecting the POS tags and removing the special characters. We filtered out the user mentions, retweet symbol (*RT*) and URLs from the text during tokenization and did not consider them as tokens. A list of words expressing feelings was obtained from *wefeelfine.org*. Twitter related slang words were obtained from a publicly available document published by United States FBI[11]. A final list of slang words was compiled by adding some more internet slangs. The list would be made available on request. Retweet counts, favorite counts, verification information, user followers count and time information were obtained from the metadata attached with each tweet returned by Twitter API. The URLs shared in tweets are generally shortened. Due to the use of different URL shortener services, a single URL might be represented in different forms by each service. In order to solve this problem, we used AlchemyAPI[12] to expand the URLs to their original form.

6.3 Baselines

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

Centroid is one of the techniques that was previously used in the literature for solving a part of our problem that ranks tweets. In order to implement it as a baseline we considered the tweets for the event in the given time period as one cluster. After preprocessing the tweets, we calculated the centroid of the cluster and ordered the tweets in the decreasing order of their similarities with the centroid. We selected *LexRank* and *TextRank*, as they are graph-based techniques for ranking textual documents, and are widely used for generating summaries. We used the open-source implementation of *LexRank* available with sumy[13] package. In case of *TextRank*, we modified the algorithm in order to make it suitable for our context. Apart from creating heterogeneous relationships in *TwitterEventInfoGraph* we also created homogeneous relationships between the *information units* as well as the tweets. Cosine similarity (≥ 0.10) was used as the measure of relatedness between tweets, and the association scores of the hashtags, text units, users and URLs were based on their co-occurrence normalized between 0 and 1. The users were associated whenever they mentioned each other in the tweets, and the association score was measured by the number of mentions normalized between 0 and 1.

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

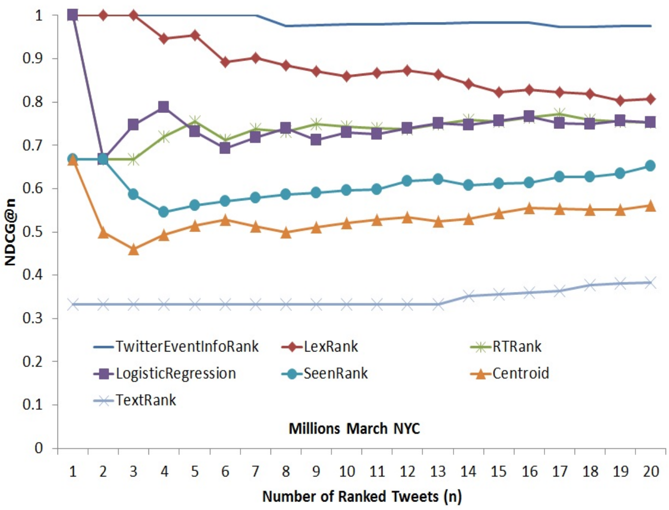

Figure 2: Performance comparison of ranking techniques using NDCG scores.

6.4 Evaluation

6.4.1 Evaluation Setup and Objectives

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

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

Tweet Annotations: The ranked tweets were annotated on an event-specific informativeness-scale of 1 to 3 by the three independent annotators. We provide sample tweets for each of them taking the Sydney

[9] 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

[10] http://nltk.org

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

[12] http://alchemyapi.com

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

Table 7: Avg IIC scores and total avg scores for annotations.

Millions March NYC	Avg 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

Sydney Siege	Avg 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

Table 8: NDCG@n and Precision@n scores for different techniques applied for ranking tweets related to Millions March NYC.

Event Name: Millions March NYC		**Recall Levels** (Note: Values for Precision are expressed in percentage)									
		@10	**@20**	**@30**	**@40**	**@50**	**@60**	**@70**	**@80**	**@90**	**@100**
TwitterEventInfoRank	*NDCG*	0.979	0.975	0.966	0.966	0.957	0.936	0.950	0.959	0.967	0.989
	Precision	100	100	100	97.5	98	96.7	95.7	95	95.5	96
LexRank	*NDCG*	0.859	0.807	0.830	0.813	0.822	0.824	0.834	0.878	0.921	0.944
	Precision	80	85	76.6	72.5	76	78.3	72.8	73.7	73.3	74
RTRank	*NDCG*	0.744	0.752	0.749	0.765	0.792	0.821	0.860	0.870	0.884	0.922
	Precision	60	70	76.6	75	70	71.6	71.4	75	73.3	69
Logistic Regression	*NDCG*	0.729	0.753	0.757	0.752	0.757	0.776	0.792	0.839	0.877	0.915
	Precision	100	100	100	97.5	96	91.6	92.8	93.7	93.3	92
SeenRank	*NDCG*	0.595	0.652	0.708	0.733	0.745	0.759	0.801	0.827	0.858	0.883
	Precision	70	65	70	67.5	62	61.6	57.1	57.5	55.5	55
Centroid	*NDCG*	0.519	0.560	0.623	0.658	0.690	0.727	0.747	0.787	0.834	0.857
	Precision	70	75	76.7	82.5	78	71.6	65.7	66.3	66.7	66
TextRank	*NDCG*	0.333	0.383	0.418	0.468	0.499	0.564	0.633	0.681	0.728	0.782
	Precision	10	5	13.3	15	22	21.6	24.3	21.3	22.2	21

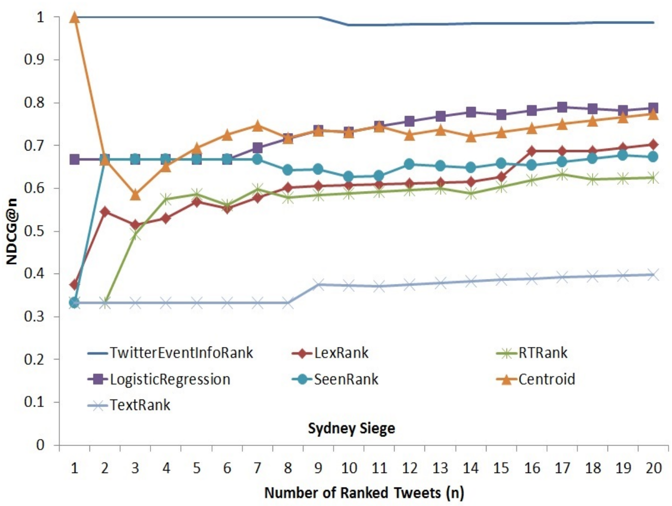

Figure 3: Performance comparison of ranking techniques using NDCG scores.

Siege event as our example. The value of 1 was assigned to tweets that does not contain any event related information (for e.g. *SteveSmith becomes Australias 45th Test captain http://t.co/nYh9DqRXxh #sydneysiege #MartinPlace Lindt #MYEFO #siege Ray Hadley Muslims ISIS*). Value of 2 was assigned to tweets that were related to the event yet they did not provide useful event-specific information (for e.g. *RT @TheDavidStevens: It wasn't just the policeman grabbing that girl in his arms, it was every Australian watching on too #sydneysiege*). A value of 3 was assigned to tweets that not only provided useful event-specific informative content but also led the user to more detailed information following the URLs mentioned in the tweet (for e.g. *RT @FoxNews: MORE: Police confirm 3 hostages escape Sydney cafe, unknown number remain inside http://t.co/pcAt91LIdS #Sydneysiege*). The annotators assigned scores to top 100 tweets ranked according to each of the seven strategies. Thereafter, we computed *Inter Indexer Consistency* (IIC) values [26] for the annotations of the two datasets. The average IIC scores obtained are shown in Table 7. The IIC values for both the events fall in the acceptable range of accuracy of annotations. A tweet might be assigned three different scores by the annotators. In that scenario we find the average of the three scores and round it off to the smallest positive integer and assign a single score to each tweet. We also report the total average scores for top 100 tweets for both the events in Table 7.

Hashtags, Text Units and URL Annotations: A similar annotation strategy was taken for annotating the top 50 hashtags, text units and URLs obtained using TwitterEventInfoRank. For hashtags and text units the annotators were asked to look at the tweets that consisted them. If the tweets primarily led to event-specific informative content then a score of 3 was assigned. If the tweets led to related but not so informative content about the event then they were assigned a score of 2. Hashtags and text units that were irrelevant and did not lead to any event related content, were assigned a score of 1. Similarly, the annotators visited the links for each URL, and based on the content they assigned them a score between 1-3. If the URLs were videos and images, then they further visited the tweet containing them in order to understand the context and scored them accordingly. Table 7 shows their average IIC scores and total average scores for top 50 ranks.

User Annotations: For annotating users we selected 5 random tweets for each of the top 50 users ranked according to *TwitterEventInfoRank*. An user was assigned a score of 3 if more than three of his tweets out of five got a score of 3 in the event-specific informativeness scale as already explained earlier. If three of his tweets get a score of 3 then the user gets a score of 2. Otherwise, a score of 1 is assigned to the user. Table 7 shows average IIC scores and total average scores for top 50 users.

NDCG@n and Precision@n: After being assured about consistency and accuracy of annotations, we moved to compute the *Normalized Discounted Cumulative Gain* (NDCG) [12] and Precision [2] values at each of the hundred recall levels. The NDCG values consider both the position and event-specific informativeness scores of the tweets. The NDCG value up-to position p in the ranking is given by equation 10, where DCG_p denotes the *discounted cumulative gain up-to position* p and is calculated using equation 9, and $IDCG_p$ denotes the *ideal discounted cumulative gain* value till position p in the ranking, or in other words the maximum possible DCG_p value till position p. rel_i denotes the graded relevance of the result at position i. In the context of our evaluation rel_i represents the average rounded score in the scale of (1-3) that has been assigned by the annotators to the tweet at position i in the ranked list of top 100 tweets.

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(i+1)} \tag{9}$$

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{10}$$

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

$$\frac{No.\ of\ relevant\ tweets\ at\ position\ n}{n} \tag{11}$$

NDCG@n and Precision@n values were calculated for all the seven approaches for each of the datasets. Figures 2 and 3 shows the NDCG curves for all the seven approaches on the Millions March NYC and the Sydney Siege events, respectively, for up-to 20 recall levels. Tables 9 and 8 presents the NDCG@n values and Precision@n values for

Table 9: NDCG@n and Precision@n scores for different techniques applied for ranking tweets related to Sydney Siege.

Event Name: Sydney Siege		**Recall Levels** (Note: Values for Precision are expressed in percentage)									
		@10	**@20**	**@30**	**@40**	**@50**	**@60**	**@70**	**@80**	**@90**	**@100**
TwitterEventInfoRank	*NDCG*	0.980	0.987	0.968	0.957	0.954	0.940	0.945	0.951	0.959	0.989
	Precision	100	100	100	100	100	100	100	97.5	96.6	96
LexRank	*NDCG*	0.607	0.701	0.684	0.707	0.736	0.768	0.763	0.805	0.838	0.867
	Precision	90	80	76.6	65	64	63.3	60	62.5	64.4	64
RTRank	*NDCG*	0.588	0.624	0.677	0.715	0.728	0.751	0.768	0.821	0.863	0.879
	Precision	80	85	86.6	85	86	88.3	90	91.3	92.22	90
Logistic Regression	*NDCG*	0.730	0.787	0.790	0.791	0.793	0.821	0.855	0.883	0.895	0.926
	Precision	60	75	76.6	72.5	74	71.6	68.5	71.3	71.1	73
SeenRank	*NDCG*	0.626	0.673	0.728	0.751	0.745	0.778	0.806	0.839	0.868	0.891
	Precision	80	85	80	75	72	68.3	70	67.5	65.5	64
Centroid	*NDCG*	0.731	0.773	0.779	0.809	0.800	0.778	0.787	0.839	0.880	0.917
	Precision	60	60	60	62.5	64	66.6	67.1	67.5	70	68
TextRank	*NDCG*	0.373	0.398	0.485	0.539	0.623	0.663	0.714	0.728	0.763	0.782
	Precision	0	10	13.3	25	28	35	42.86	45	47.8	51

Table 10: Top five event-specific informative hashtags, text units, URLs and tweets for Sydney Siege

Event	**Sydney Siege**
Top 5 Event-specific Informative Hashtags	1. #sydneysiege, 2. #SydneySiege, 3. #Sydneysiege, 4. #MartinPlace, 5. #9News
Top 5 Event-specific Informative Text Units	1. police, 2. sydney, 3. reporter, 4. lindt, 5. isis
Top 5 Event-specific Informative Urls	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 5 Event-specific Informative Tweets	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

Table 11: Three Randomly Selected Tweets for Top Three Event-specific Informative Users of Sydney Siege.

Three Randomly Selected Tweets for Top Three Event-specific Informative Users
User 1. 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 #sydneysiâĂę
User 2. 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 #SydneyâĂŹs 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. 1.,RT @RT_com: #SYDNEY: Gunman tall man in late 40s, dressed in black âĂŞ 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://âĂę

different recall levels upto 100. It is quite evident from the figures and the tables that TwitterEventInfoRank approach outperforms all the baselines including the state-of-the-art approach of *SeenRank* in gaining event-specific information.

6.5 Discussion

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

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

We also show the top 5 event-specific informative hashtags, text units, URLs and tweets for the Sydney Siege event in Table 10, and three randomly selected tweets for top three event-specific informative users for the same event in Table 11. Due to space constraints we do not present similar tables for Millions March NYC.

7. CONCLUSION AND FUTURE WORK

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

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

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

8. REFERENCES

[1] Alejandro, M., and Paloma, M. The use of metrics for measuring informality levels in web 2.0 texts.

[2] Baeza-Yates, R., Ribeiro-Neto, B., et al. *Modern information retrieval*, vol. 463. ACM press New York, 1999.

[3] Barrons, G. Śsuleiman: Mubarak decided to step down# egypt# jan25 oh my godŠ: Examining the use of social media in the 2011 egyptian revolution. *Contemporary Arab Affairs 5*, 1 (2012), 54–67.

[4] Becker, H., Naaman, M., and Gravano, L. Beyond trending topics: Real-world event identification on twitter. *ICWSM 11* (2011), 438–441.

[5] Becker, H., Naaman, M., and Gravano, L. Selecting quality twitter content for events. *ICWSM 11* (2011).

[6] Damak, F., Pinel-Sauvagnat, K., Boughanem, M., and Cabanac, G. Effectiveness of state-of-the-art features for microblog search. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing* (2013), ACM, pp. 914–919.

[7] Duan, Y., Jiang, L., Qin, T., Zhou, M., and Shum, H.-Y. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics* (2010), Association for Computational Linguistics, pp. 295–303.

[8] Erkan, G., and Radev, D. R. Lexrank: graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* (2004), 457–479.

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

[10] Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd international conference on World Wide Web companion* (2013), International World Wide Web Conferences Steering Committee, pp. 1021–1024.

[11] Inouye, D., and Kalita, J. K. 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* (2011), IEEE, pp. 298–306.

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

[13] Laniado, D., and Mika, P. Making sense of twitter. In *The Semantic Web–ISWC 2010*. Springer, 2010, pp. 470–485.

[14] Lin, J., and Schatz, M. Design patterns for efficient graph algorithms in mapreduce. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs* (2010), ACM, pp. 78–85.

[15] McCreadie, R., and Macdonald, C. Relevance in microblogs: Enhancing tweet retrieval using hyperlinked documents. In *Proceedings of the 10th conference on open research areas in information retrieval* (2013), LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, pp. 189–196.

[16] Mihalcea, R., and Tarau, P. Textrank: Bringing order into texts. Association for Computational Linguistics.

[17] Nagmoti, R., Teredesai, A., and De Cock, M. Ranking approaches for microblog search. In *2010 IEEE/WIC/ACM International joint conference on Web Intelligence-Intelligent Agent Technology (WI-IAT)* (2010), vol. 1, IEEE Computer Society, pp. 153–157.

[18] Nichols, J., Mahmud, J., and Drews, C. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces* (2012), ACM, pp. 189–198.

[19] O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM 11* (2010), 122–129.

[20] Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. 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)* (2014), no. EPFL-CONF-203561.

[21] Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation ranking: Bringing order to the web.

[22] Phelan, O., McCarthy, K., and Smyth, B. Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems* (2009), ACM, pp. 385–388.

[23] Popescu, A.-M., Pennacchiotti, M., and Paranjpe, D. Extracting events and event descriptions from twitter. In *Proceedings of the 20th international conference companion on World wide web* (2011), ACM, pp. 105–106.

[24] Purohit, H., and Sheth, A. P. Twitris v3: From citizen sensing to analysis, coordination and action. In *ICWSM* (2013).

[25] Ritter, A., Clark, S., Etzioni, O., et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2011), Association for Computational Linguistics, pp. 1524–1534.

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

[27] Sakaki, T., Okazaki, M., and Matsuo, Y. Tweet analysis for real-time event detection and earthquake reporting system development. *Knowledge and Data Engineering, IEEE Transactions on 25*, 4 (2013), 919–931.

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

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

[30] Vosecky, J., Leung, K. W.-T., and Ng, W. Searching for quality microblog posts: Filtering and ranking based on content analysis and implicit links. In *Database Systems for Advanced Applications* (2012), Springer, pp. 397–413.

[31] Wei, F., Li, W., Lu, Q., and He, Y. 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* (2008), ACM, pp. 283–290.

[32] Weng, J., Lim, E.-P., Jiang, J., and He, Q. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining* (2010), ACM, pp. 261–270.

[33] Yardi, S., Romero, D., Schoenebeck, G., et al. Detecting spam in a twitter network. *First Monday 15*, 1 (2009).