Identification and Ranking of Event-specific Entity-centric Informative Content from Twitter

Debanjan Mahata[1], John R. Talburt[1], and Vivek Kumar Singh[2]

[1] University of Arkansas at Little Rock, Department of Information Science, USA
[2] South Asian University, Department of Computer Science, New Delhi, India
dxmahata@ualr.edu, jrtalburt@ualr.edu, vivek@cs.sau.ac.in

Abstract. Twitter has become the leading platform for mining information related to real-life events. A large amount of the shared content in Twitter are non-informative spams and informal personal updates. Thus, it is necessary to identify and rank informative event-specific content from Twitter. Moreover, tweets containing information about named entities (like person, place, organization, etc) occurring in the context of an event, generates interest and aids in gaining useful insights. In this paper, we develop a novel generic model based on the principle of mutual reinforcement, for representing and identifying event-specific, as well as entity-centric informative content from Twitter. An algorithm is proposed that ranks tweets in terms of event-specific, entity-centric information content by leveraging the semantics of relationships between different units of the model.

1 Introduction

Twitter is a social media 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. Twitter is widely accepted as a source for first-hand citizen journalistic content and has been harnessed in detection, extraction and analysis of real-life events [4,3].

Motivation: A significant amount of tweets are related to real-life events (e.g, football matches, music shows, etc). Majority of these event related tweets are pointless babbles, personal updates and spams providing no information to the general audience interested to know about an event. On the other hand there are tweets that presents newsworthy content, recent updates and real-time coverage of on-going events. These tweets are informative and are very useful for users who follow an event, and search for related information in Twitter.

Occurrence of a real-life event in general is characterized by participation of entities like people, organizations, or things at a certain place over a period of time [5]. While sharing information about an event in Twitter, users often mention these entities (e.g *Update: Statement from Australian Prime Minister Tony Abbott on the Hostage incident #SydneySiege http://t.co/b4tO4A8CQj*). We consider such user updates as entity-centric messages related to the event. The consumers of event related information are most often interested in such entity-centric messages in the context of the event. Also, informative content shared about the entities during an event helps in gaining useful insights about the event as well as the related entities.

Objective and Contribution: The main objective of the work presented in this paper is to automatically identify and rank event-specific informative tweets mentioning relevant entities in their content. Towards this objective, we propose a novel generic model based on principle of mutual reinforcement for representing relationships between event-specific information cues and relevant named entities extracted from the tweet content. We implement an algorithm that leverages the mutually reinforcing relationships represented by the model for ranking tweets in terms of event-specific informative content sharing information about entities related to the event.

Problem: Events have been defined from various perspectives and in different contexts. In the context of our work we adopt a definition similar to [1]. An **event** is defined as a real-world occurrence (E_i) with an associated time period T_{E_i} ($t_{E_i}^{start}$-$t_{E_i}^{end}$), and a time ordered stream of tweets M_{E_i}, of substantial volume, discussing the occurrence of the event and posted in time T_{E_i}. We formally state the problem as,

Given an event E_i, a time ordered stream of n tweets $M_{E_i}=\{m_1,m_2,...,m_n\}$ related to the event posted in time period T_{E_i}, the problem is to find a ranked set of tweets $\hat{M}_{E_i}=\{m_1 \geq ... \geq m_i \geq m_j \geq ... \geq m_n | i<j\}$, ordered in decreasing order of its event-specific informative content sharing information about event related entities.

2 Methodology

Content of a tweet is primarily composed of hashtags, words for expressing and conveying information, and URLs that lead to additional information about the content. While conveying information about an event the users also mention named entities in the textual content of the tweets. The example tweet in the previous section, not only provides information about the Sydney Siege crisis event, but also informs about “Tony Abbott” in the context of the event. The tweets are posted by users. It is also intuitive that users having high follower count tends to post informative posts, as tweets posted by such users are read by a larger audience and vice versa. Therefore, for an event E_i, in order to identify and rank event-specific informative tweets discussing about entities relevant to the event, we consider the following as event-specific information units:

- set of *tweets* ($M_{E_i}=\{m_1,m_2,...,m_n\}$) related to the event.
- set of *hashtags* ($H_{E_i}=\{h_1,h_2,...,h_p\}$) used for annotating the event related tweets.
- set of *entities* ($W_{E_i}=\{w_1,w_2,...,w_r\}$) mentioned in the event related tweets.
- set of *users* ($U_{E_i}=\{u_1,u_2,...,u_s\}$) posting tweets ($\in M_{E_i}$) about the event.
- set of *URLs* ($L_{E_i}=\{l_1,l_2,...,l_t\}$) linking to external sources related to the event

The informativeness w.r.t an event for any information unit depends upon its occurrence with other information units. We define event-specific informativeness of each unit based on the assumption below. For an event E_i

- a *tweet is considered to be event-specific informative* if it is strongly associated with: **(a)** *event-specific informative hashtags*, **(b)** *event-specific informative entities*, **(c)** *event-specific informative users*, **(d)** *event-specific informative URLs*.

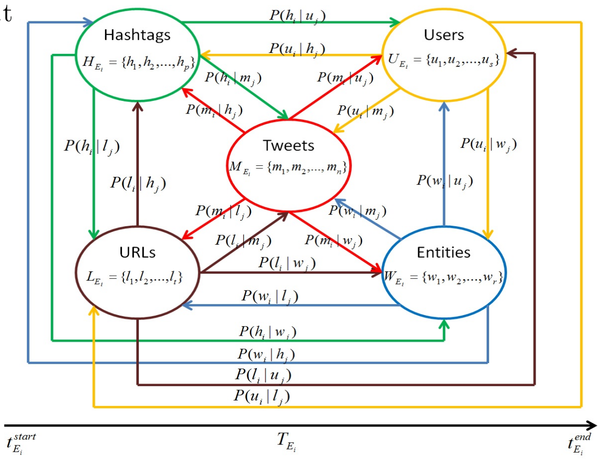

Fig. 1. Mutual Reinforcement Chains in Twitter.

We similarly define event-specific hashtags, entities, users and URLs forming a circular mutually reinforcing relationships between each other as shown in Fig 1.

The relationships between the event-specific *information units* for an E_i forms a *Mutual Reinforcement Chain* [6], as shown in Fig 1. We represent this relationship in a graph $G=(V,D)$, where $V=M_{E_i} \cup H_{E_i} \cup W_{E_i} \cup U_{E_i} \cup L_{E_i}$, is the set of vertices and 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 1. We do not consider an edge between two vertices of same type.

We assign an initial event-specific score to all the vertices of the graph. The formulations of the scores assigned to the vertices $\in H_{E_i}, W_{E_i}, U_{E_i}, L_{E_i}$ can be found in Table 1. For initializing the tweets $(\in M_{E_i})$ with a generic informativeness score we develop a logistic regression model with an accuracy of 76.32% after 10-fold cross validation. For training the model we used an annotated dataset provided by [2]. 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. The 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. The assigned initial scores gives an initial ranking of the vertices. We aim to refine the initial scores and assign a final score for ranking the vertices by leveraging the relationships between them and propagating the initial scores accordingly, from one vertex to another. Next, we formalize our ranking methodology and present our proposed algorithm step-by-step.

The relationships between two sets of vertices in the graph G is denoted by an affinity matrix. For example, $A_{E_i}^{MH}$ denotes the $M_{E_i}-H_{E_i}$ affinity matrix for event E_i, where $(i,j)^{th}$ entry is the edge weight quantifying the association between i^{th}

Table 1. Equations for mutual reinforcement chains, affinity scores and event-specific initialization scores of nodes $\in G$.

Affinity scores between different nodes $\in M_{E_i}, H_{E_i}, W_{E_i}, U_{E_i}, L_{E_i}$:

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

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

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

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

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

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

$$P(h_i|m_j)=P(m_i|h_j)=P(w_i|m_j)=P(m_i|w_j)=P(u_i|m_j)=P(m_i|u_j)=P(l_i|m_j)=P(m_i|l_j)=1.0$$

Note: $P(h_i \mid w_j)$ should be read as the probability of occurrence of hashtag h_i given the occurrence of the entity 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 nodes $\in H_{E_i}, W_{E_i}, U_{E_i}, L_{E_i}$:

$$Score(h_i)=\frac{freq(h_i)}{max\{freq(h_1),freq(h_2),...,freq(h_p)\}} \quad \textbf{(1)} \qquad Score(w_i)=\frac{freq(w_i)}{max\{freq(w_1),freq(w_2),...,freq(w_r)\}} \quad \textbf{(2)}$$

$$Score(u_i)=\frac{followers(u_i)}{max\{followers(u_1),...,followers(u_r)\}} \quad \textbf{(3)}, \qquad Score(l_i)=\frac{freq(l_i)}{max\{freq(l_1),freq(l_2),...,freq(l_r)\}} \quad \textbf{(4)}$$

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} entity ($\in W_{E_i}$) 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 $u_i \in (U_{E_i})$.

Equations representing mutual reinforcement chains between $M_{E_i}, H_{E_i}, W_{E_i}, U_{E_i}, L_{E_i}$:

$$R_{E_i}^{M(k+1)} = A_{E_i}^{MM(k)} R_{E_i}^{M(k)} + A_{E_i}^{MH(k)} R_{E_i}^{H(k)} + A_{E_i}^{MW(k)} R_{E_i}^{W(k)} + A_{E_i}^{MU(k)} R_{E_i}^{U(k)} + A_{E_i}^{ML(k)} R_{E_i}^{L(k)} \quad \textbf{(5)}$$

$$R_{E_i}^{H(k+1)} = A_{E_i}^{HM(k)} R_{E_i}^{M(k)} + A_{E_i}^{HH(k)} R_{E_i}^{H(k)} + A_{E_i}^{HW(k)} R_{E_i}^{W(k)} + A_{E_i}^{HU(k)} R_{E_i}^{U(k)} + A_{E_i}^{HL(k)} R_{E_i}^{L(k)} \quad \textbf{(6)}$$

$$R_{E_i}^{W(k+1)} = A_{E_i}^{WM(k)} R_{E_i}^{M(k)} + A_{E_i}^{WH(k)} R_{E_i}^{H(k)} + A_{E_i}^{WW(k)} R_{E_i}^{W(k)} + A_{E_i}^{WU(k)} R_{E_i}^{U(k)} + A_{E_i}^{WL(k)} R_{E_i}^{L(k)} \quad \textbf{(7)}$$

$$R_{E_i}^{U(k+1)} = A_{E_i}^{UM(k)} R_{E_i}^{M(k)} + A_{E_i}^{UH(k)} R_{E_i}^{H(k)} + A_{E_i}^{UW(k)} R_{E_i}^{W(k)} + A_{E_i}^{UU(k)} R_{E_i}^{U(k)} + A_{E_i}^{UL(k)} R_{E_i}^{L(k)} \quad \textbf{(8)}$$

$$R_{E_i}^{L(k+1)} = A_{E_i}^{LM(k)} R_{E_i}^{M(k)} + A_{E_i}^{LH(k)} R_{E_i}^{H(k)} + A_{E_i}^{LW(k)} R_{E_i}^{W(k)} + A_{E_i}^{LU(k)} R_{E_i}^{U(k)} + A_{E_i}^{LL(k)} R_{E_i}^{L(k)} \quad \textbf{(9)}$$

Other equations:

$$\Delta_{E_i}.R_{E_i} = \lambda.R_{E_i} \quad \textbf{(10)} \qquad \overline{\Delta}_{E_i} = \alpha\hat{\Delta}_{E_i} + (1-\alpha)E \quad \textbf{(11)} \qquad E = p \times [1]_{1\times k} \quad \textbf{(12)}$$

tweet ($\in M_{E_i}$) and j^{th} hashtag ($\in H_{E_i}$), calculated using Table 1, and so on. The rankings of tweets, hashtags, entities, 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 M_E, H_{E_i}, W_{E_i}, U_{E_i}, and L_{E_i}, respectively. Therefore, the Mutual Reinforcement Chain ranking for the k^{th} iteration can be formulated using equations (5-9) in Table 1. 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

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

Input : Sets of vertices $M_{E_i}, H_{E_i}, W_{E_i}, U_{E_i}, L_{E_i}$ of graph G, $\alpha = 0.85$, $\varepsilon = 1e-08$.
Output: Ordered set of vertices $\hat{M}_{E_i}$, containing tweets ranked in order of event-specific informative content sharing information about event related entities.

Initialize rank vectors $[R_{E_i}^{M(0)}, R_{E_i}^{H(0)}, R_{E_i}^{W(0)}, R_{E_i}^{U(0)}, R_{E_i}^{L(0)}]$;
Assign $R_{E_i}^{0} = [R_{E_i}^{M(0)}, R_{E_i}^{H(0)}, R_{E_i}^{W(0)}, R_{E_i}^{U(0)}, R_{E_i}^{L(0)}]^T$;
Normalize $R_{E_i}^{0}$ such that $||R_{E_i}^{0}||_1=1$;
Construct matrix Δ_{E_i};
Make matrix Δ_{E_i} stochastic and irreducible converting it to $\overline{\Delta}_{E_i}$;
$k \leftarrow 1$
repeat
  $R_{E_i}^{k} \leftarrow \overline{\Delta}_{E_i} R_{E_i}^{k-1}$;
  $k \leftarrow k+1$;
until $||\ R_{E_i}^{k} - R_{E_i}^{k-1}\ ||_1 < \varepsilon\ OR\ k \geq 100$;
$R_{E_i}^{M} \leftarrow R_{E_i}^{M(k)}, R_{E_i}^{H} \leftarrow R_{E_i}^{H(k)}, R_{E_i}^{W} \leftarrow R_{E_i}^{W(k)}, R_{E_i}^{U} \leftarrow R_{E_i}^{U(k)}, R_{E_i}^{L} \leftarrow R_{E_i}^{L(k)}$;
$\hat{M}_{E_i} \leftarrow R_{E_i}^{M}$;
return $\hat{M}_{E_i}$

Fig. 2. Algorithm for ranking nodes of graph G.

then, R_{E_i} can be computed as the dominant eigenvector of Δ_{E_i}, as shown in equation 10 in Table 1. 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 $\overline{\Delta}_{E_i}$ using equations 11 and 12 in Table 1, where $0 \leq \alpha \leq 1$ is set to 0.85, and k is the order of $\hat{\Delta}_{E_i}$. We set $p = [1/k]_{k \times 1}$ by assuming a uniform distribution over all elements. Now, $\overline{\Delta}_{E_i}$ is stochastic and irreducible and it can be shown that it is also primitive by checking $\overline{\Delta}_{E_i}^2$ is greater than 0. The final ordered set of tweets $\hat{M}_{E_i}$ ranked in terms of their event-specific informative content sharing information about entities related to the event is obtained using the algorithm in Fig 2.

3 Results and Future Work

Table 2. 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

For implementing our proposed framework 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 2. Tweets for each event was collected over the given period of time, by providing a popular hashtag corresponding to each event to the API. We only considered English language tweets. We performed a series of data preparation steps before implementing the logistic regression model and our algorithm. Some of the steps that we took are deduplication of tweets, tokenization, POS tagging, detection of slang words, English stop words, feeling words, and special characters [3]. We extracted named entities from the tweets using AlchemyAPI (http://alchemyapi.com). The entities containing slang words were removed. Removal of slang hashtags and entities

[3] list of resources like slang words, stopwords and feeling words used can be obtained from https://github.com/dxmahata/EIIMFramework/tree/master/CodeBase/EventIdentityInformationManagement/Resources

was done in order to obtain high quality results as intuitively high quality informative tweets should not contain a lot of slangs.

Given the space constraint, we show the top 5 hashtags, entities URLs and tweets for the Sydney Siege event in Table 3. We do not report the users for privacy concerns. Apart from identifying event-specific informative tweets containing information about event related entities, our proposed model has an additional advantage of identifying and ranking top event-specific informative hashtags, entities, URLs and users for an event. In this paper we presented our methodology and some preliminary results. Our next step would be to evaluate our results rigorously and extend the developed framework in a distributed computing environment.

Table 3. Top 5 informative hashtags, entities, URLs and tweets for Sydney Siege

Event	**Sydney Siege**
Top 5 Informative Hashtags	1. #sydneysiege, 2. #SydneySiege, 3. #Sydneysiege, 4. #MartinPlace, 5. #9News
Top 5 Informative Entities	1. police, 2. sydney, 3. reporter, 4. lindt, 5. isis
Top 5 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 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, 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

References

1. H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. *ICWSM*, 11:438–441, 2011.
2. A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM" 14)*, number EPFL-CONF-203561, 2014.
3. A.-M. Popescu, M. Pennacchiotti, and D. Paranjpe. Extracting events and event descriptions from twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 105–106. ACM, 2011.
4. T. Sakaki, M. Okazaki, and Y. Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *Knowledge and Data Engineering, IEEE Transactions on*, 25(4):919–931, 2013.
5. R. Shaw, R. Troncy, and L. Hardman. Lode: Linking open descriptions of events. In *The Semantic Web*, pages 153–167. Springer, 2009.
6. F. Wei, W. Li, Q. Lu, and Y. He. Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 283–290. ACM, 2008.