Identifying Event-Specific Sources from Social Media

Debanjan Mahata and Nitin Agarwal

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

Abstract. Social media has become an indispensable resource for coordinating various real-life events by providing a platform to instantly tap into a huge audience. The participatory nature of social media creates an environment highly conducive for people to share information, voice their opinion, and engage in discussions. It is not uncommon to find novel and specific information with intimate details for an event on social media platforms in contrast to the mainstream media. This makes social media a valuable source for event analysis studies. It is, therefore, of utmost importance to identify quality sources from these social media sites for understanding and exploring an event. However, due to the power law distribution of the Internet, social media sources get buried in the Long Tail. The overwhelming number of social media sources makes it even more challenging to identify the valuable sources. We propose an evolutionary mutual reinforcement model for identifying and ranking highly 'specific' social media sources and 'close' entities related to an event. Due to the absence of ground truth, we provide a novel evaluation strategy for validating the model. By considering the top ranked sources according to our model, we observe a substantial information gain (ranging between 25% and 130%) as compared to the baselines (viz., Google search and Icerocket blog search). Moreover, highly informative sources are ranked much higher according to our model as compared to the widely-used baselines, putting spotlight on the social media sources that could be easily overlooked otherwise. Our model further affords an apparatus to analyze events at micro and macro scales. Data for the research is collected from various blogging platforms such as, Blogger (hosted at blogspot), LiveJournal, WordPress, Typepad, etc. and will be made publicly available for researchers.

Keywords: event analysis, social media, blogs, mutual reinforcement, specificity, closeness, information gain

1 Introduction

Social media has brought a paradigm shift in the way people share information and communicate. Social media played an important role in mobilizing events such as, 'The Arab Spring', 'Occupy Wall Street', 'Sandy relief efforts', 'London

Riots','The Spanish Revolution', among others. This led to a surge in citizen journalism all over the world, encouraging transnational participation. Thus, social media serves as a parallel, yet distinct source of information about real-life events along with the mainstream media space [32].

The mainstream media sources often gloss over the intricate details while covering a real-life event. The information could be biased, regulated by the government, and may not present a well-rounded report of an event [15]. On the contrary, social media sources often contain uninhibited and unedited opinions of the masses. Blogs, especially, are widely accepted in the blogging community as sources of more holistic information with intricate details of an event when compared to mainstream media sources [19]. Thus the sources, which are obtained from social media could potentially provide a rather 'closer' or an on-the-ground view of the events with novel information. The information gleaned from social media affords opportunities to study various social phenomenon from methodological and theoretical perspectives including, situation awareness for better crisis response, humanitarian assistance and disaster relief, social movements, citizen and participatory journalism, collective action [2–4], and more.

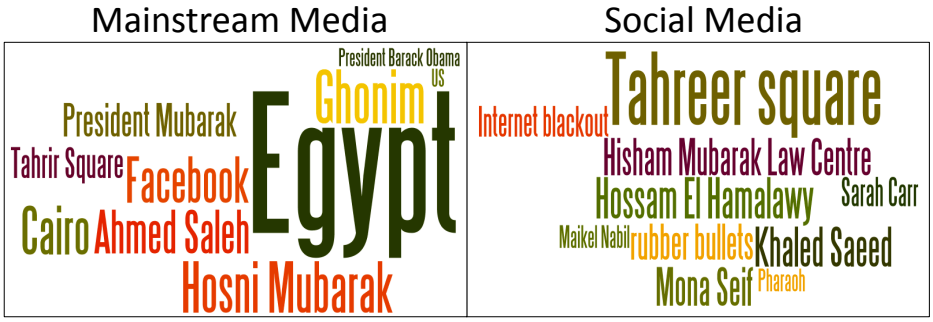

Fig. 1. Top 10 entities from mainstream media and blogs.

Motivation: An initial analysis of the top 10 entities obtained from the top 10 search results related to "Egyptian Revolution" from two mainstream media channels (BBC and CNN), and from blogs during the time of the revolution is shown in Fig 1. The top entities from the mainstream media channels are certainly relevant but fairly broad level, meaning they do not contribute specific or intricate details about the revolution. In contrast, the top entities from the blogs provide intimate details about the events associated with the revolution. For example, the activists like 'Mona Seif', 'Sarah Carr', 'Maikel Nabil' and 'Hosam El Hamalwy' were very closely involved, and were responsible for mobilizing the event. The entities like 'Internet Blackout' and 'Khaleed Saeed' were central to the event. Moreover, the presence of entities like 'Facebook' and 'Ghonim' (who coordinated the event on Facebook) among the top mainstream media entities also indicates the significance of social media in the event.

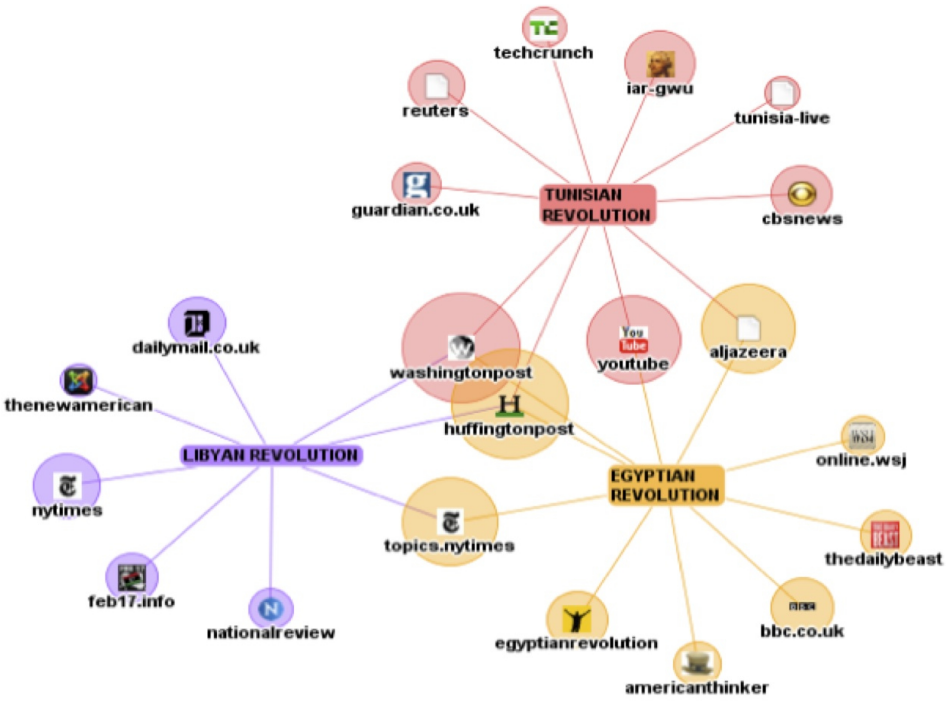

Fig. 2. Top 10 Google search results for "Egyptian Revolution", "Libyan Revolution", and "Tunisian Revolution", visualized using TouchGraph.

Due to the power law distribution of the Internet [1], and the present search engine technology, the top search results, or the 'Short Head', is generally dominated by the mainstream media websites. As illustrated in Fig 2 the top 10 search results for "Egyptian Revolution", "Libyan Revolution", and "Tunisian Revolution" returned by Google, visualized using Touchgraph[1], retrieved mainstream media sources. Consequently, the social media sites get buried in the "Long Tail" [25] of the search result distribution as shown in Fig 3. However sources from the social media channels, act as hubs of specific information about real-life events [16]. Thus, a person interested to analyze an event may miss out the novel and specific information available in social media by relying on the top results from the popular search engines. Moreover, in the words of Chris Anderson [6], *"With an estimated 15 million bloggers out there, the odds that a few will have something important and insightful to say are good and getting better."* This motivated us to look for techniques in this chapter, that would help in identifying these otherwise buried sources providing highly specific information related to an event.

[1] http://touchgraph.com

Challenges: Identifying highly informative 'specific' sources and 'close' entities related to a real-life event from social media entails various challenges as follows,

- **Sparsity of sources:** Enormous population of the sparsely linked Long Tail social media sources
- **Quality assessment dilemma:** The entities (person, organization, place, etc.) mentioned in the sources act as the atomic units of information. Sources which are 'specific' to an event must contain entities 'closer' or highly relevant to the event. On the other hand, such 'close' entities can be obtained from the 'specific' sources. This presents a dilemma in assessing the quality of the sources for event related 'specific' information content, and makes it a nontrivial task.
- **Entity extraction:** It is also a challenge, to accurately extract the entities from the social media sources, which are mostly unstructured and have colloquial content.
- **Lack of evaluation measures:** Conventional information retrieval based evaluation measures help in identifying the most relevant and authoritative sources, however, these sources may not be the most novel or offer specific information. Therefore, new evaluation measures are required to estimate the performance of our work.

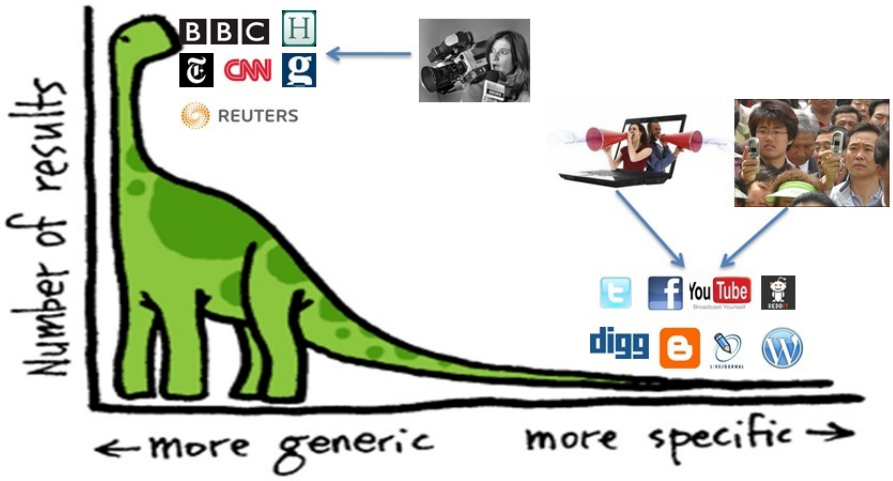

Fig. 3. Short Head Vs Long Tail media sources.

Contributions: We make the following contributions,

- **Methodology:** A methodology based on the principle of mutual reinforcement, that helps in identifying highly 'specific' sources and 'close' entities

from social media, and their relationships (Section 4.2). It ranks the sources and the entities based on their 'specific' information content and how 'close' they are with respect to a set of events.
- **Evaluation Strategy:** Present the methodology, along with an objective evaluation strategy to validate our findings (Section 6.3).
- **Experiment on sources related to real-life events:** Perform our experiments on sources and entities related to the events: 'Egyptian Revolution', 'Libyan Revolution', and 'Tunisian Revolution' (Section 6). However, the work is extendible to other types of events.
- **Event analysis:** Explore the utility of such a model in analyzing events (Section 7) and conclude the work with future directions (Section 8).

Next, we present the related work and compare and contrast these with the proposed approach, highlighting our contributions to the literature.

2 Related Work

Due to huge number of informal sources in social media it is a difficult task to identify high quality sources related to real-life events. Researchers have built semantic web models for efficient retrieval of event related media sources [36]. Event related contents have been found leveraging the tagging and location information associated with the photos shared in Flickr [31]. *Becker et al* [7], studied how to identify events and high quality sources related to them from Twitter. In order to identify the genuine sources of information, credibility and trustworthiness of event related information were studied from Twitter [14]. New methods were investigated for filtering and assessing the verity of sources obtained from social media for journalists [10]. All these works, try to explore the quality of information, in terms of relevancy, usefulness, timeliness of the content and usage patterns of authoritative users producing the content. Moreover, none of them involves the blogosphere. However, our work investigates on specific information content in blogs related to a real-life event by using the named entities that are closely associated with the event. The specificity scores of the blogs help in quickly gaining novel and specific information about an event as shown in Section 6.3. The method improves the quality of event-specific information gained by users from the ranked sources.

Several methods have been developed in the past for identifying and ranking quality sources from the web [5]. PageRank [9] took advantage of the link structure of the web for ranking web pages. It was further improved for making it sensitive to topic based search [17]. Graph based approaches were used for modeling documents and a set of documents as weighted text graphs, and for computing relative importance of textual units for Natural Language Processing [12]. Mutual reinforcement principle was used for identifying Hubs and Authorities from a subset of web pages using HITS algorithm [21]. The main idea of our algorithm is similar to that of the HITS algorithm. Instead of finding highly authoritative web pages and hubs we find specific sources and entities in the

context of an event. Moreover, we propose an evolutionary model that demonstrates faster convergence and better performance as shown in Section 6.2. The same principle has been used to solve the problem of identifying reliable users and content from social media [8, 20], as well as tracking discovered topics in web videos [24]. To our knowledge there is no work that explores relationships between named entities and sources from social media for reinforcing the identification of specific information using the Mutual Reinforcement principle.

User-generated data from various social media platforms, related to real-life events, have been studied to perform wide range of analysis. Platforms like TwitterStand [34], Twitris [18], TwitInfo [28] and TweetXplorer [29] have developed techniques to provide analytics, and visualizations related to different real-life events. Similar tools have been used for tracking earthquakes [33], providing humanitarian aid during the time of crisis [22], analyzing political campaigns [37] to studying socio-political events [35]. Most of the event analysis frameworks rely on finding relevant keywords and networks between the content producers in order to analyze events. Our work primarily deals with named entities for extracting event-specific information. However, what makes it different from all the other event analysis frameworks is its capability to distinguish highly specific entities from the generic ones among the relevant entities for an event. The entities thus identified helps in further analyzing the event from different perspectives as explained in Section 7.

3 Problem Definition

The number of sources related to an event in social media is overwhelming. All these sources may not provide useful information and needs to be processed in order to identify the valuable sources providing specific information about the concerned event. Provided we have a set of events, a set of sources, and a set of entities related to each of these events, we need to rank these sources and entities from the most specific to the most generic ones, based on their information content.

Event: *We define an event to be a real-world incident, occurring at any place at any time or over a certain period of time.*

Specificity and Closeness: *Given a finite set of events ξ, we take an event $E_j \in \xi$ such that, $1 \leq j \leq \mid \xi \mid$, a set of 'p' sources denoted by ϕ_{E_j}, and a set of 'q' entities denoted by σ_{E_j}, related to the event E_j. We define two functions κ (specificity) and τ (closeness) such that:*

$$\kappa : S_i \rightarrow [0, 1] \tag{1}$$

$$\tau : e_i \rightarrow [0, 1] \tag{2}$$

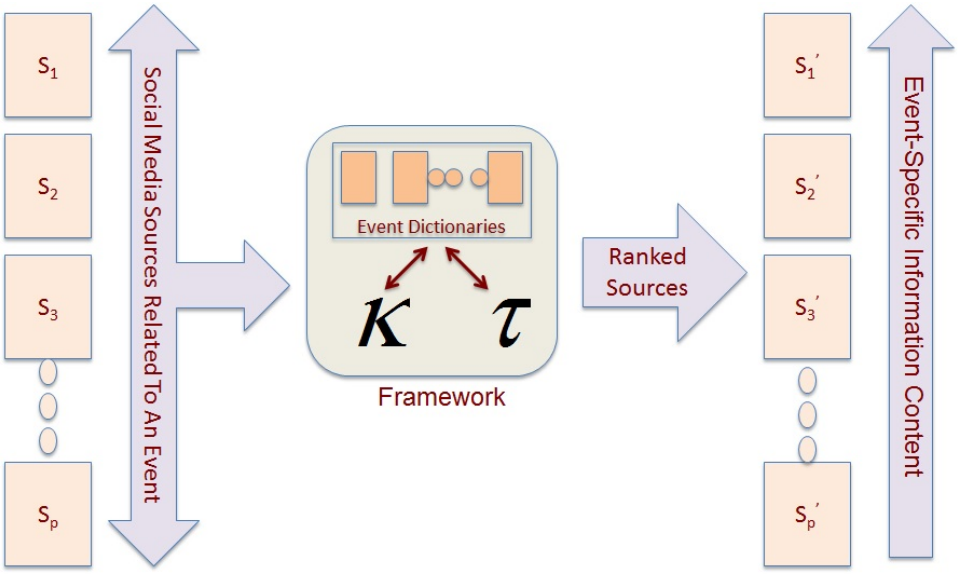

Fig. 4. Black box view of the problem.

where, $S_i(\in \phi_{E_j})$*, is the* i^{th} *source, and* $e_i(\in \sigma_{E_j})$ *is the* i^{th} *entity, so that we can get two ordered sets* (φ_{E_j} *and* ς_{E_j}) *for the set of sources in* ϕ_{E_j} *and entities in* σ_{E_j}*, such that:*

$$\varphi_{E_j} = \{S_1, ..., S_i, S_j, ..., S_p \mid \kappa(S_i) \geq \kappa(S_j), i < j\} \tag{3}$$

$$\varsigma_{E_j} = \{e_1, ..., e_i, e_j, ..., e_q \mid \tau(e_i) \geq \tau(e_j), i < j\} \tag{4}$$

φ_{E_j} is ordered in decreasing order of how 'specific' S_i is w.r.t E_j. ς_{E_j} is ordered in decreasing order of how 'close' e_i is w.r.t E_j. A black-box view of the problem is shown in Fig 4.

4 Methodology

A real-life event is characterized by a distinct set of close entities (persons, places, organizations, etc.) along with generic ones. The entities act as the basic units of information in these sources as shown in Fig 5. Intuitively, specific sources would contain closer entities and one is likely to find closer entities in more specific sources. The relation between specific sources and close entities could then be modeled following the Mutual Reinforcement Principle, which forms the basis of our methodology. The methodology presented here discusses a more rigorous treatment of the problem over our previous studies [27, 26].

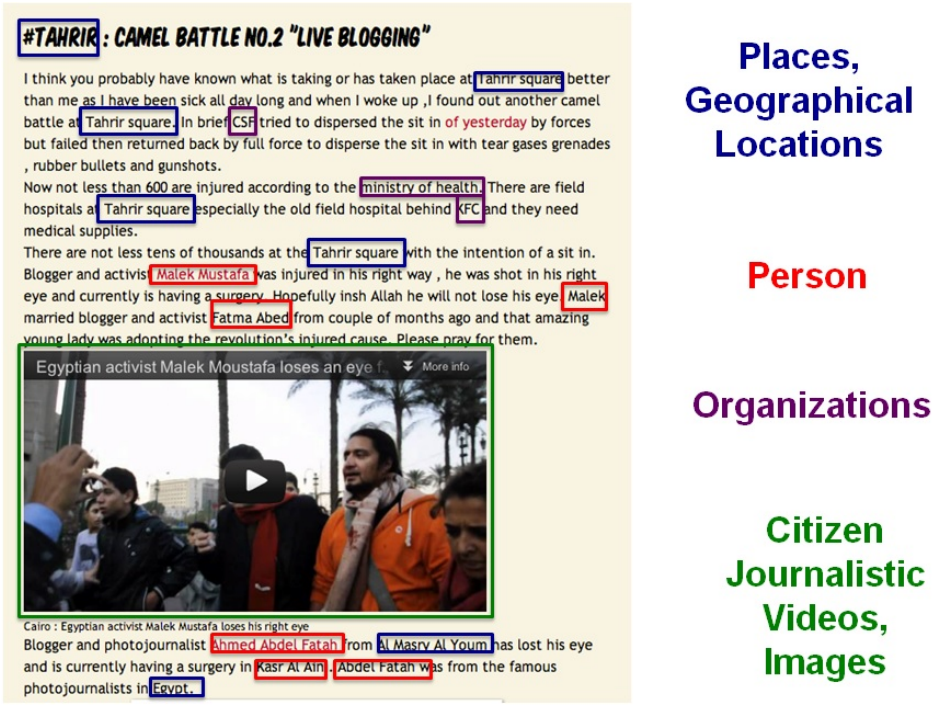

Fig. 5. Entities associated with a social media source.

> An entity should have high 'closeness' score if it appears in many sources with high 'specificity' scores while a source should have a high 'specificity' score if it contains many entities with high 'closeness' scores.

In essence the principle states that the 'closeness' score of an entity is determined by the 'specificity' scores of the sources it appears in, and the 'specificity' score of a source is determined by the 'closeness' scores of the entities it contain. The proposed methodology extends the basic Mutual Reinforcement Principle to consider the evolving knowledge learned about an event. However, the model requires an apriori or seed knowledge about an event, which is provided in terms of an event profile or an event dictionary. Next, we discuss the construction of event dictionaries.

4.1 Event Dictionaries

Each event E_j is profiled by constructing an event dictionary (σ_{E_j}). In order to calculate specificity of a source w.r.t an event, we need to start with an initial set of close entities. At the same time, these close entities are better acquired from the specific sources. To solve this dilemma, we construct event dictionaries, from independent sources which are completely separate from the sources (ϕ_{E_j}) that need to be ranked.

Formulation of initial closeness scores: We calculate the 'closeness' score $(\tau(e_i)_{E_j})$ of each entity (e_i) for event E_j in order to construct the event dictionaries, by using equations 5 and 6 based on tf-idf measure [30] , from the information retrieval literature. Let $E_j \in \xi$, be the j^{th} event, and 'e_i' be the i^{th} entity extracted from the set of sources selected for constructing the event dictionaries. If the term $f(e_i, E_j)$ denotes the frequency of occurrence of the entity 'e_i' in the set of sources for the event E_j, and $IE_jf(e_i)$ denotes the inverse event frequency for the entity 'e_i' then closeness score $(\tau(e_i)_{E_j})$ of an entity e_i w.r.t the event E_j is defined as,

$$\tau(e_i)_{E_j} = e_i f\_IE_j f = f(e_i, E_j) * IE_j f(e_i) \tag{5}$$

$$IE_j f(e_i) = log(\frac{\mid \xi \mid}{\mid E_j \in \xi : e_i \in E_j \mid}) \tag{6}$$

and, $\mid E_j \in \xi : e_i \in E_j \mid$ refers to the number of events in which the entity e_i occurs. Since we extract the entities from the sources related to the events, we cannot have an entity that does not belong to any of the events. Therefore, we always have $\mid E_j \in \xi : e_i \in E_j \mid > 0$.

We get $\mid \xi \mid$ number of event dictionaries, each corresponding to an event. Following steps are taken to construct the event dictionaries:

1. ***Entity Extraction:*** Entities are extracted from all the sources collected from GlobalVoices[2] as explained in Section 5, using AlchemyAPI[3] and their corresponding $\tau(e_i)_{E_j}$ values are calculated using equation 5. We choose GlobalVoices for obtaining the seed sources for constructing the initial event dictionaries, as it is a portal where bloggers and translators work together to make reports of various real-life events, from blogs and citizen media everywhere.This makes it a reliable source for finding specific information content from social media. Due to colloquial nature of the sources as discussed in the challenges, some of the entities occur in several forms. For example, the entity 'Tahrir Square' occur as 'Tahreer', 'El-Tahrir', etc. We resolve such multiple representation of the same entity by applying pattern matching[4]. Given two entities represented as strings we accept them to be the same if their patterns match by 80% or more. We would like to use the standard entity resolution algorithms in our future work.
2. ***Closeness Score Computation:*** For each event E_j, we calculate $\tau(e_i)_{E_j}$ scores for the set of entities for that event using equations 5 and 6. An entity may occur in multiple events and hence can be present in multiple event dictionaries with different $\tau(e_i)_{E_j}$ scores.
3. ***Ranking:*** The higher the $\tau(e_i)_{E_j}$ score of an entity the closer it is to the event. The entities are then ranked according to the descending $\tau(e_i)_{E_j}$ scores.

[2] http://globalvoicesonline.org
[3] http://alchemyapi.com
[4] http://docs.python.org/2/library/difflib.html

4. ***Normalization:*** Since the range of closeness scores are different for each event, we normalize $\tau(e_i)_{E_j}$ scores w.r.t an event between 0 and 1. The normalization enables an assessment of relative closeness of an entity across multiple events.

The dictionaries thus obtained from the above mentioned procedure are static and serve as a good source of apriori knowledge about the event. However, as we discover new knowledge from specific sources, it is desirable to update the event dictionaries. However, the method applied for constructing the initial event dictionaries require a set of events. This is a drawback of the current method and we plan to improve it in a future work. Next, we discuss how the dictionaries help in identifying specific sources, which in turn help in improving the dictionary.

4.2 Mutually Reinforcing Sources and Entities

Given an event $E_j \in \xi$, a set of sources (ϕ_{E_j}) and entities (σ_{E_j}), related to the event, we define two column vectors: ***'Specificity'*** ($\kappa_{\mathbf{E_j}}$) and ***'Closeness'*** ($\tau_{\mathbf{E_j}}$).

$$\kappa_{\mathbf{E_j}} =< \kappa(S_1)_{E_j}, \kappa(S_2)_{E_j}, ..., \kappa(S_p)_{E_j} >^T \tag{7}$$

$$\tau_{\mathbf{E_j}} =< \tau(e_1)_{E_j}, \tau(e_2)_{E_j}, ..., \tau(e_q)_{E_j} >^T \tag{8}$$

where, $\kappa(S_i)_{E_j}$ ($\in range(\kappa)$, from equation 1) represents the *'specificity'* score of i^{th} source $S_i (\in \phi_{E_j})$, for $1 \leq i \leq p$ and $\tau(e_i)_{E_j}$ ($\in range(\tau)$, from equation 2) represents the *'closeness'* score of i^{th} entity $e_i (\in \sigma_{E_j})$, for $1 \leq i \leq q$. Each source S_i may contain related as well as unrelated information about various events. If we consider the set of events ξ, then each $\kappa(S_i)_{E_j}$ is itself a vector of 'specificity' values of the source S_i w.r.t the events ($\in E_j$) as expressed in equation 9.

$$\kappa(S_i)_{E_j} =< \kappa(S_i)_{E_1}, \kappa(S_i)_{E_2}, ..., \kappa(S_i)_{E_{|\xi|}} > \tag{9}$$

Similarly, each entity e_i may be related to various events. If we consider the set of events ξ, then each $\tau(e_i)_{E_j}$ is itself a vector of 'closeness' values of the entity e_i w.r.t the events ($\in E_j$) as expressed in equation 10.

$$\tau(e_i)_{E_j} =< \tau(e_i)_{E_1}, \tau(e_i)_{E_2}, ..., \tau(e_i)_{E_{|\xi|}} > \tag{10}$$

However, while representing $\kappa(S_i)_{E_j}$ and $\tau(e_i)_{E_j}$ as an element of the vectors $\kappa_{\mathbf{E_j}}$ and $\tau_{\mathbf{E_j}}$, respectively, we only choose the entry for the j^{th} event under consideration.

We construct a bipartite graph $G = (V, U)$ (Figure 6) representing the mutual relationship between the sources and the entities, where $V \in \phi_{E_j}, \sigma_{E_j}$, is the set of vertices for G, and U is the set of undirected edges. The sources without entities are discarded during this process.

The presence of an entity in a source is not sufficient to determine its specificity. In order to express the specificity of a source w.r.t an event, we need

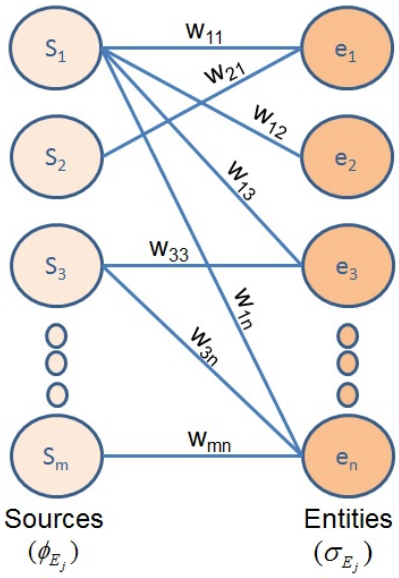

Fig. 6. Bipartite graph G representing the mutual relationship between the sources and the entities.

to consider the closeness value of the entities present in the source. Given the closeness $\tau(e_n)_{E_j}$ of an entity e_n w.r.t an event E_j, obtained from the event dictionary, a weight w_{mn} is assigned to the edges of the graph, which expresses the magnitude by which an entity is related to a source S_m.

The significance of an entity e_n in a source S_m is expressed as,

$$\frac{f(e_n, S_m)}{\sum_{n=0}^{q} f(e_n, S_m)} \tag{11}$$

where, $f(e_n, S_m)$ is the frequency of occurrence of the entity e_n in the source S_m. Therefore mathematically,

$$w_{mn} = \frac{\tau(e_n)_{E_j} * f(e_n, S_m)}{\sum_{n=0}^{q} f(e_n, S_m)} \tag{12}$$

The adjacency matrix of the bipartite graph G is denoted by L, and is defined as follows:

$$L_{mn} = \begin{cases} w_{mn} & if\ (m,n) \in U \\ 0 & otherwise \end{cases}$$

Input: Set of sources ϕ_{E_j}, set of entities σ_{E_j} from the event dictionary, threshold for convergence μ.
Output: Ordered set of sources φ_{E_j}, ranked according to their 'specificity' scores w.r.t E_j, and, ordered set of entities ς_{E_j}, ranked according to their 'closeness' scores w.r.t E_j.

Initialize $\kappa_{\mathbf{E_{j(0)}}} \leftarrow \tau_{\mathbf{E_{j(0)}}} \leftarrow < 1, 1, ..., 1 >$;
Initialize $k \leftarrow 1$;
repeat
  Construct matrices $\mathbf{L_{k-1}}$ and $\mathbf{L^T_{k-1}}$;
  Calculate matrix products $M \leftarrow \mathbf{L_{k-1}L^T_{k-1}}$ and $M'' \leftarrow \mathbf{L^T_{k-1}L_{k-1}}$;
  Convert M and M'' into stochastic matrices $M_{stochastic} \leftarrow M$ *and* $M''_{stochastic} \leftarrow M''$;
  $\kappa_{\mathbf{E_j}}(\mathbf{k}) \leftarrow M_{stochastic}\kappa_{\mathbf{E_j}}(\mathbf{k-1})$;
  $\tau_{\mathbf{E_j}}(\mathbf{k}) \leftarrow M''_{stochastic}\tau_{\mathbf{E_j}}(\mathbf{k-1})$;
  normalize $\kappa_{\mathbf{E_j}}(\mathbf{k}) \leftarrow \frac{\kappa_{\mathbf{E_j}}(\mathbf{k})}{||\kappa_{\mathbf{E_j}}(\mathbf{k})||_1}$;
  normalize $\tau_{\mathbf{E_j}}(\mathbf{k}) \leftarrow \frac{\tau_{\mathbf{E_j}}(\mathbf{k})}{||\tau_{\mathbf{E_j}}(\mathbf{k})||_1}$;
  $k \leftarrow k + 1$;
  Update σ_{E_j} ;
until $|| \kappa_{\mathbf{E_j}}(\mathbf{k}) - \kappa_{\mathbf{E_j}}(\mathbf{k-1}) ||_1 < \mu$ *and* $|| \tau_{\mathbf{E_j}}(\mathbf{k}) - \tau_{\mathbf{E_j}}(\mathbf{k-1}) ||_1 < \mu$;
Reverse sort $\kappa_{\mathbf{E_j}}(\mathbf{k})$,$\tau_{\mathbf{E_j}}(\mathbf{k})$;
return $\kappa_{\mathbf{E_j}}(\mathbf{k})$,$\tau_{\mathbf{E_j}}(\mathbf{k})$;

Fig. 7. Algorithm for calculating 'specificity' and 'closeness'.

Following the Mutual Reinforcement Principle the relationships between specificity scores of sources and closeness scores of entities for event E_j can be denoted as follows,

$$\kappa_{\mathbf{E_j}} = \mathbf{L}\tau_{\mathbf{E_j}} \tag{13}$$

$$\tau_{\mathbf{E_j}} = \mathbf{L^T}\kappa_{\mathbf{E_j}} \tag{14}$$

Substituting the values for $\kappa_{\mathbf{E_j}}$ and $\tau_{\mathbf{E_j}}$, we derive the following equations,

$$\kappa_{\mathbf{E_j}} = \mathbf{LL^T}\kappa_{\mathbf{E_j}} \tag{15}$$

$$\tau_{\mathbf{E_j}} = \mathbf{L^TL}\tau_{\mathbf{E_j}} \tag{16}$$

Equations 15 and 16 are characteristic equations of an eigensystem, where the solutions to $\kappa_{\mathbf{E_j}}$ and $\tau_{\mathbf{E_j}}$ are the respective eigen vectors with the corresponding eigenvalue of 1.

To emphasize the relationship between the sources and the entities, we make a major contribution by modifying the way the equations 15 and 16 are solved. We make the matrices $\mathbf{LL^T}$ and $\mathbf{L^TL}$ evolutionary while solving the equations. Since each of the equations is a circular definition, the final specificity and closeness scores are computed using the power iteration method [13]. Each iteration improves specificity and closeness scores reflecting their mutual relationship. As we move towards getting the specific sources and close entities in each iteration, we update the weights (w_{mn}) assigned to the edges between the sources and the entities by the newly calculated closeness scores for the entities. This results in renewed reinforcement of the relationship at every iteration by getting closer entities from better sources and vice-versa. This essentially helps the model incorporate the newly discovered knowledge about the events. More precisely, the improved understanding of the relationship between the source and the entities vis-a-vis an event is incorporated into the model.

The updation of the edge weights and the matrices with k^{th} iteration is represented as follows,

$$w_{mn(k)} = \frac{\tau(e_{n(\mathbf{k-1})})_{E_j} * f(e_n, S_m)}{\sum_{n=0}^{q} f(e_n, S_m)} \tag{17}$$

$$L_{mn(k)} = \begin{cases} w_{mn(k)} & if\ (m,n) \in U \\ 0 & otherwise \end{cases}$$

where, $L_{mn(k)}$ represents the adjacency matrix for graph G, and $w_{mn(k)}$ denotes the edge weight for the edge between m^{th} source and n^{th} entity at the k^{th} iteration. $\tau(e_{n(\mathbf{k-1})})_{E_j}$ represents the closeness score of the entity e_n w.r.t the event $E_j (\in \xi)$, obtained from the evolving event dictionary for event E_j at $(k-1)^{th}$ iteration.

If, $\kappa_{\mathbf{E_j}}(\mathbf{k})$ and $\tau_{\mathbf{E_j}}(\mathbf{k})$ be the specificity and the closeness scores, at the k^{th} iteration, the iterative process for generating the final solution are,

$$\kappa_{\mathbf{E_j}}(\mathbf{k}) = \mathbf{L_{k-1}L_{k-1}^T}\kappa_{\mathbf{E_j}}(\mathbf{k-1}) \tag{18}$$

$$\tau_{\mathbf{E_j}}(\mathbf{k}) = \mathbf{L_{k-1}^TL_{k-1}}\tau_{\mathbf{E_j}}(\mathbf{k-1}) \tag{19}$$

In order to get 1 as the largest eigenvalue and, $\kappa_{\mathbf{E_j}}$ and $\tau_{\mathbf{E_j}}$ as the principal eigen vectors, the matrices $\mathbf{L_{k-1}L_{k-1}^T}$ and $\mathbf{L_{k-1}^TL_{k-1}}$ needs to be stochastic and irreducible [23] at every step of our evolutionary process. In the present case, since the graph G is a bipartite graph, matrices $\mathbf{L_{k-1}L_{k-1}^T}$ and $\mathbf{L_{k-1}^TL_{k-1}}$ are already irreducible.

In order to make the matrices $\mathbf{L_{k-1}L_{k-1}^T}$ and $\mathbf{L_{k-1}^TL_{k-1}}$ stochastic, we take the following steps at each iteration,

- Dividing the non-zero entries of the matrices $\mathbf{L_{k-1}L_{k-1}^T}$ and $\mathbf{L_{k-1}^T L_{k-1}}$ by the summation of all the entries in a row.
- Assigning $1/n$ to the zero entries of $\mathbf{L_{k-1}L_{k-1}^T}$ and $1/m$ to the zero entries of $\mathbf{L_{k-1}^T L_{k-1}}$, respectively.

The whole process is presented as an algorithm as shown in Fig 7.
We also perform our study using conventional binary static matrices represented as follows,

$$L_{mn} = \begin{cases} 1 & if\ (m,n) \in U \\ \\ 0 & otherwise \end{cases}$$

The proposed evolutionary model outperforms the static model as validated by the results discussed in Section 6.

5 Data Collection

Motivation behind source selection: For many people blogs have become popular social media sources for satisfying interpersonal communication needs. Blogs act as a platform for masses to share their likes and dislikes, voice their opinions, provide suggestions and report news. Over the years blogging has matured from personal diaries to citizen journalistic sources providing live coverage of events beyond the professional newsrooms. Often mainstream media rely on blogs for reporting first-hand accounts of an event [11]. Other social media platforms like microblogs, social networks etc., also promote such activities. But, these platforms have very little scope to elaborately discuss about the events due to the limitations in the length of content allowed to be posted. However, these alternative platforms act as good sources for studying and tracking dissemination of information during real-life events. This motivated us to perform our experiments on sources collected from the blogging platforms instead of other social media websites.

Table 1. Details of Data Collected.

Service Used	Event	Number of Blog Posts
GlobalVoices	Egyptian Revolution	234
	Libyan Revolution	86
	Tunisian Revolution	77
Google Blogger	Egyptian Revolution	579
	Libyan Revolution	600
	Tunisian Revolution	484
Icerocket Blog Search	Egyptian Revolution	5900
	Libyan Revolution	2198
	Tunisian Revolution	1220

Sources: Blog posts from GlobalVoices, Blogger[5] and Icerocket Blog Search[6] respectively, are collected for the study. The details of the dataset used is given in Table 1. The dataset includes 11,378 blog posts from various blogging platforms like blogspot.com, wordpress.com, livejournal.com, typepad.com, etc. We also filter out the non-English blogs. The data from GlobalVoices is used for constructing event dictionaries (σ_{E_j}), as explained in Section 4. We collect blog posts related to the three events from Blogger using Google search, and from other blogging platforms using Icerocket blog search. We perform our experiments on the sources (ϕ_{E_j}) retrieved by the search engines due to the lack of ground truth and take the sources along with the ranks assigned to them by the search engines as our baseline (explained in Subsection 6.3) The collected blog posts are parsed for extracting various information. However, we use the following information for our study: *URL* of the blog and blogpost), *blog text*, *entities*, *language*, and *rank* of the post in the respective search engines used for collecting it. We use AlchemyAPI in order to extract entities. These datasets would be made available on request.

6 Experiment and Analysis

In this section, we describe the experiments performed. First, we discuss the experimental setup, followed by the comparative analysis between the proposed evolutionary mutual reinforcement model and conventional mutual reinforcement model. We then, introduce a novel evaluation strategy comparing the proposed model with two baseline models.

6.1 Experimental Setup

The methodology discussed earlier is implemented on the collected datasets. We take the following steps in order to perform the experiment,

- ***Constructing the Event Dictionaries:*** We take $\xi = \{$"Egyptian Revolution", "Libyan Revolution", "Tunisian Revolution"$\}$ as our set of events. We construct the event dictionaries (σ_{E_j}) by using the sources from GlobalVoices (explained in Event Dictionary subsection).
- ***Implementing the Proposed Evolutionary Mutual Reinforcement Model:*** The algorithm, as presented in Figure 7, is implemented on the set of sources (ϕ_{E_j}) from Blogger and Icerocket related to each event respectively, and the set of entities (σ_{E_j}) from the event dictionary corresponding to each event E_j. The threshold value for convergence μ is set to 1e-08.
- ***Obtaining Specific Sources and Close Entities:*** With the termination of the algorithm we get a ranked set of sources (φ_{E_j}) ordered in terms of their specific information content and entities (ς_{E_j}) ordered in terms of how closely related they are to the event.

[5] http://blogger.com
[6] http://icerocket.com

– ***Conventional Mutual Reinforcement Model Approach:*** We also implement the conventional mutual reinforcement model on ϕ_{E_j} and σ_{E_j} without considering the evolving matrices (explained in Subsection 4.2).

6.2 Comparing Conventional and Evolutionary Mutual Reinforcement Models

In order to compare the efficiency of the proposed evolutionary mutual reinforcement model with the conventional mutual reinforcement model we combine the sources collected for an event E_j from Google search and Icerocket search. After that we run the proposed evolutionary mutual reinforcement model and conventional mutual reinforcement model on the combined set of sources and event dictionary (σ_{E_j}) for each event. The number of iterations taken by the power iteration method to converge in each case is analyzed in Figure 8. It is observed that the number of iterations taken by the power iteration method is lesser, when we employ the evolutionary mutual reinforcement model. Hence, we observe a marked improvement in the performance of our evolutionary mutual reinforcement model over the conventional static mutual reinforcement model.

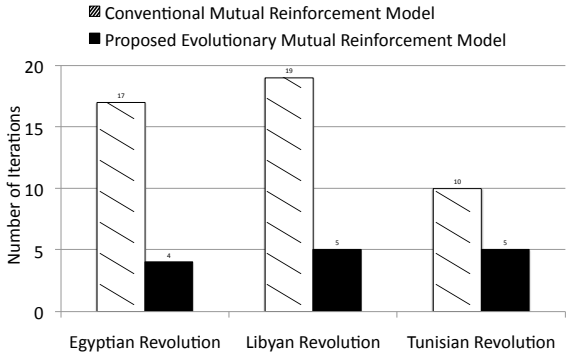

Fig. 8. Comparison of number of iterations taken by the power iteration method to converge, for the set of sources related to events in ξ, with the proposed evolutionary mutual reinforcement model and the conventional static mutual reinforcement model.

Explanation for the improvement: The lesser number of iterations in the proposed evolutionary mutual reinforcement model can be explained by the introduction of the evolutionary weights assigned to the relationship between the sources and the entities. The improved closeness scores ($\tau(e_i, E_j)$) at each iteration of the evolutionary model and the renewed reinforcement of the relationship between the entities and the sources decreases the number of iterations taken by the proposed evolutionary model to converge in comparison to the static conventional model. Next, we compare the performance of the proposed evolutionary model with the baselines and the conventional mutual reinforcement model in

terms of how quickly the models help in identifying valuable information about the events.

Egyptian Revolution

Specificity Based Ranking	Google Search Ranking
1	59
2	286
3	400
4	277
5	55
6	202
7	6
8	9
9	313
10	374

Libyan Revolution

Specificity Based Ranking	Google Search Ranking
1	13
2	329
3	9
4	194
5	24
6	311
7	364
8	204
9	374
10	184

Tunisian Revolution

Specificity Based Ranking	Google Search Ranking
1	162
2	40
3	420
4	459
5	72
6	181
7	152
8	440
9	99
10	174

Egyptian Revolution

Specificity Based Ranking	Icerocket Blog Search Ranking
1	75216
2	10607
3	53924
4	56604
5	9831
6	25790
7	1
8	99925
9	94614
10	53924

Libyan Revolution

Specificity Based Ranking	Icerocket Blog Search Ranking
1	47276
2	11751
3	4900
4	22501
5	4
6	11040
7	43520
8	11751
9	41631
10	18271

Tunisian Revolution

Specificity Based Ranking	Icerocket Blog Search Ranking
1	9713
2	42985
3	36335
4	3843
5	46784
6	42645
7	99
8	1
9	63141
10	42645

Fig. 9. Rankings of the sources from Google Blogger and Icerocket based on 'specificity' (κ) values obtained from our model and the rankings assigned by Google Search and Icerocket Blog Search.

6.3 Baseline Comparisons

Selecting the Baselines: Due to lack of benchmark datasets we use the search results obtained from Google and Icerocket Blog Search as baselines for validation. Standard information retrieval measures for evaluation (DCG, NDCG, MAP, MAP@10) could not be used due to the absence of ground truth. We also consider the sources ranked according to the conventional static mutual reinforcement model and show the effectiveness of our model in quickly gaining information about an event. We further analyze the ranking of the top K specific sources as identified by our methodology and observe the difference in their

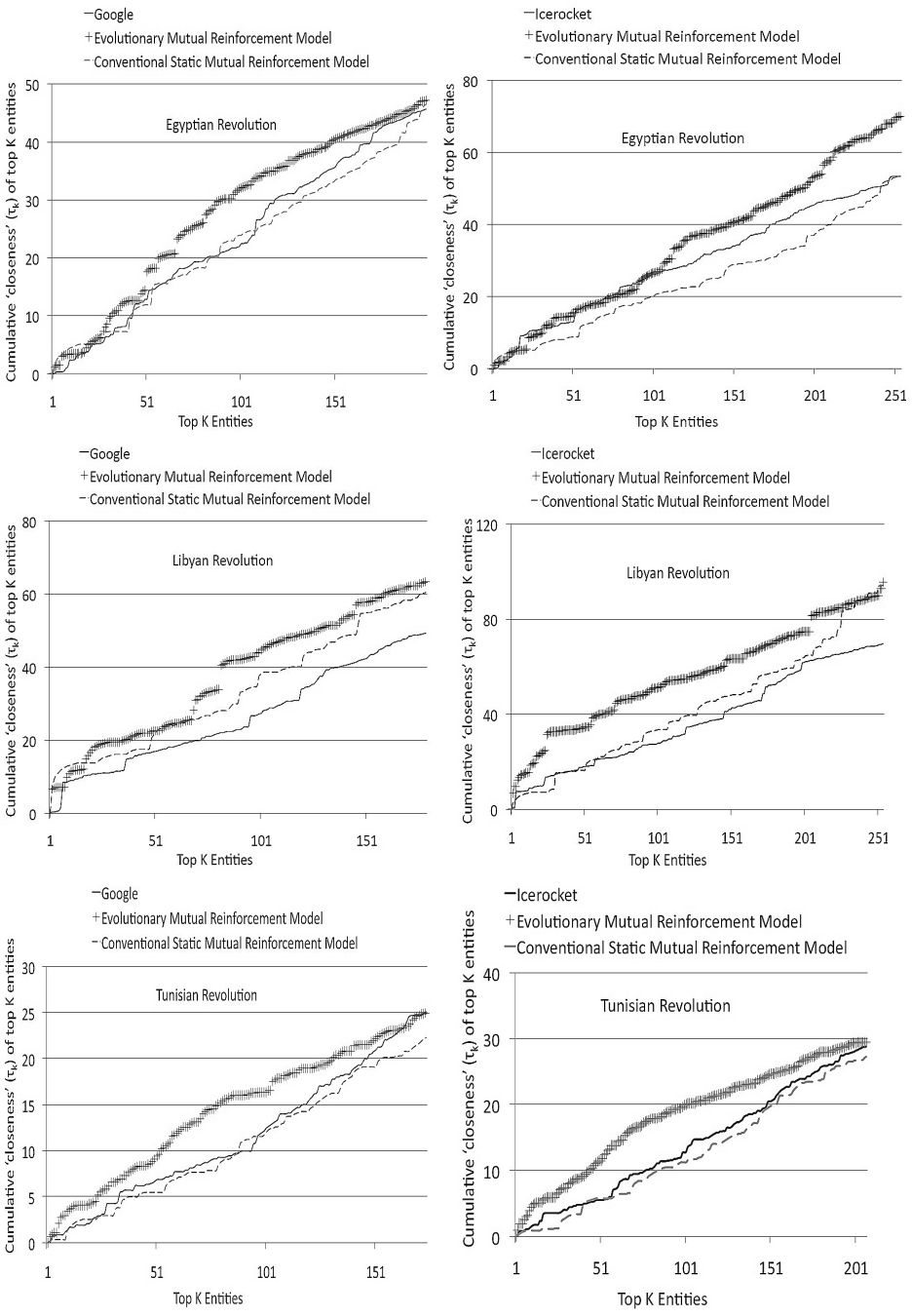

Fig. 10. Validating specific sources obtained from our model.

rankings as assigned by the search engines and as assigned by our model. Figure 9 shows ranks assigned by Google and Icerocket for the top 10 sources for each event, ranked according to our framework. We conclude, that our framework could identify the sources that often gets buried in the Long Tail and has the potential for presenting valuable information about the event. Next, we propose a novel strategy to demonstrate the effectiveness of the specific sources ranked by our evolutionary mutual reinforcement model in identifying highly informative sources.

Rationale Behind The Strategy: Search engines are designed to give the most relevant sources containing the close entities, related to a query for a given event. As these close entities have a very high probability to be associated with the event, we expect to gain valuable information about the event from these entities making the highly ranked sources very specific to the event. We use this notion to propose a novel evaluation strategy, showing that when sources related to an event are ranked according to our model, they provide more valuable information than the ranking order given by the search engines and the conventional mutual reinforcement model for the same set of sources.

Implementation of The Strategy: We compare closeness ($\tau(e_i)_{E_j}$) values of the entities obtained from the sources ranked according to the search engines, the conventional model and our model respectively. Following steps are taken,

- ***Preparing the Ranked Lists of Entities:*** From the three differently ranked lists (search engine, our evolutionary model and conventional static model), each source is visited and the entities are extracted from them. The '$\tau(e_i)_{E_j}$' values are assigned to these entities by referring to the respective final event dictionary (ς_{E_j}) and they are ranked in descending order.
- ***Measuring the Information Gain:*** As we traverse the three list of sources, we obtain three list of same entities, arranged in different orders depending upon the ranking of the sources. In order to show the comparison in the gain of information from the three lists we take the top 'K' entities from each list and calculate the sum of their '$\tau(e_i)_{E_j}$' values and plot them against the value of 'K' in Figure 10. We start from K=1 and go on increasing its value till the number of entities are exhausted in all the three lists. It is evident from Figures 10, that the curves based on specificity ($\kappa(S_i)_{E_j}$) quickly gains over the curves based on the search engines and the conventional model.
- ***Analysis:*** Figure 10 shows that information about the event is gained quicker using our model, which could identify specific sources earlier than the search engines as well as the conventional model. We measure the maximum percentage gain of information in each set of sources related to the three events. There is a maximum gain of information (130.4%) in case of sources obtained from Icerocket search related to 'Tunisian Revolution', and a minimum gain of information (25.97%) in case of the sources obtained from Icerocket search related to 'Egyptian Revolution'. Also when the sources are ranked according to our methodology they gain the maximum percentage of information at the 9^{th} source in case of sources obtained from Google

related to 'Libyan Revolution'. This in turn implies that the sources ranked higher by our model are more specific than the ones ranked by the search engines and the conventional model. These highly specific sources are also very informative. When presented earlier they also help in learning useful information about the event due to the presence of close entities in them. As we already observed earlier that these sources are often Long Tail sources, we can conclude that Long Tail sources that are ranked lower by the search engines, when identified are often more specific than the highly ranked short head sources.

Table 2. Top 5 entities in the event specific and the event class dictionaries constructed for the set of events ξ.

Egyptian Revolution Specific Dictionary	Tahrir Square, Egyptian government, Gigi Ibrahim, Alexandria, Wael Abbas.
Libyan Revolution Specific Dictionary	Tripoli, Muammar Al Gaddafi, North Atlantic Treaty Organization, Chad, United Kingdom
Tunisian Revolution Specific Dictionary	Tunisian government, Lin Ben Mhenni, Samir Feriani, Kasbah Square, RCD
Socio-Political event dictionary	Twitter, Iranian Government, Tear gas devices, Facebook, Big Social network

7 Further Exploration

We use the proposed framework developed by us as an apparatus to further explore event characteristics and show its utility in analyzing events. We show the potential of the final event dictionaries (ς_{E_j}) obtained for each event E_j for gaining valuable information about the event and to identify the generic entities related to the class of events.

The entities in the final event dictionaries (ς_{E_j}) are examined further. Based on their frequency $f(e_i, E_j)$ and closeness ($\tau(e_i)_{E_j}$) scores, we categorize the entities into the following categories: a. *Close and Frequent*, b. *Close but Not Frequent*, c. *Not Close but Frequent*, and d. *Neither Close nor Frequent* as shown in Figure 7. We categorize an entity in each event dictionary as 'frequent' if it occurs more than a threshold value η i.e $f(e_i, E_j) \geq \eta$, and 'close' if it has a closeness score more than a threshold value α i.e $\alpha \geq \tau(e_i, E_j)$. After careful manual inspection we decided the values of η and α to be 5 and 0.5 respectively. The values of η and α is same for all the events $E_j \in \xi$. However, different thresholds could be examined in the future though. Each entry in a matrix,

	Frequent (η >= 5)	(η < 5) Not Frequent
(α >= 0.5) Close	Egyptian government, Tahrir Square, Suez, Gigi Ibrahim, Alexandria, Maikel Nabil, NDP, Muslim Brotherhood, Egyptian Army, Wael Ghonim.	Internet Café, Day of Anger, Mariam Arafat, Candice Holdsworth, Dalia Al Marghani, Sherine Tadros, EGP, Bahaa El-Tawil, Mir Hussein Mousavi, Hussein Sharif.
Not Close (α < 0.5)	Anwar al-Sadat, Open Society Institute, Facebook, Professor Rashid Khalidi, dictator, Mohammed Abdel Dayem, mass movement, illegal occupation, Tea Party, Abdel Salam Karmen.	Mainstream media, Oman, Tunis, Carlos Latuff, Ukraine, Sudan, Cuba, Hillary Clinton.

Egyptian Revolution

	Frequent (η >= 5)	(η < 5) Not Frequent
(α >= 0.5) Close	Saif Al Islam Gaddaffi, Pentagon, oil, Mohamed Nabbous, Central Africa, Muammar Gaddafi, Rwanda, Arab League, Ethiopia, Khamis Gaddaffi	North Atlantic Treaty Organization, Iyad El Baghdadi, Libyan State Television, Bent Benghazi, Altarash, Benina airport, Omar Al-Mukhtar, Youssef Al Qaradawi, Eman Al Obaidi, Hamdi Kadri
Not Close (α < 0.5)	Tony Buckingham, Senoussis Brotherhood, Safia Farkash, Cynthia McKinney, United States of America, Canada, popular media, United Africa, Emad Benosman, Hosni Mubarak	Mainstream media, Oman, Tunis, Bahrain, UAE, Saudi Arabia, New York Times, Afghanistan, Ben Ali, Wael Ghonim

Libyan Revolution

	Frequent (η >= 5)	(η < 5) Not Frequent
(α >= 0.5) Close	Ennahda party, Moncef Marzouki, Sidi Bouzid, Tunisian government, Habib Bourguiba, Al Abedine Ben Ali, Muslim Brotherhood, RCD, Mohammed Ghannouchi.	Kasbah Square, France, Slim Amamou, Amira Yahyaoui, Beji Caid Al Sebsi, Mehdi Lamloum, Sami Ben Gharbia, Aziza Othmana Hospital, Samar Dahmash Jarrah, Sami Ben Romdhane
Not Close (α < 0.5)	Michele Alliot-Marie, social networks, facebook, Arabs, Jeffrey Feltman, United States of America, Al-Mahdi, Al-Qaida, Al Jazeera, Kareem Salama	Mainstream media, Oman, Doha, Tunis, UAE, Saudi Arabia, New York Times, Hosni Mubarak, Bill Clinton, Hillary Clinton

Tunisian Revolution

Fig. 11. Different categories of entities for the events.

consists of top 10 entities that satisfy the thresholds for the corresponding row and column category. We further examine these entities in the context of each of these events and report interesting observations. We present the observations for "Egyptian Revolution", next, however similar observations were made for other events.

7.1 Event-specific Popular and Close Entities

We analyze the entities in the *Close and Frequent*, and *Close but Not Frequent* categories. We find that *Close and Frequent* entities are not only frequent but are also closely related to the event. In other words these are very popular entities related to the event. For example, the occurrence of 'Tahrir Square' and 'Egyptian government' in this category, is inevitable, as Tahrir Square is the place where the protest started against the Egyptian Government. These are also the entities that anyone comes to know about from the top search results given by the search engines as well as from mainstream media sources.

On the other hand, *Close but Not Frequent* category primarily consists of entities that add novel and useful insights to the events. The occurrence of 'Internet Cafe' in this category clearly shows how the local people in Egypt used Internet Cafes for accessing various social media websites in order to coordinate and participate in the revolution. It is also the place where 'Khaled Said', a 28 year old computer-wiz was arrested by the Egyptian police. He was brutally tortured to death, triggering the anger among the people of Egypt for a mass uprising against the dictatorial government. January 25, 2011 is also known as 'The Day of Anger' in the Egyptian Revolution, as it is the day that marked the start of a series of protests and riots in Egypt. So the occurrence of the entity, 'Day of Anger' in this category is reasonable.

Drawing upon the differences between the two categories our findings suggests that although the category of *Close and Frequent* entities give a lot of information about an event, it is also necessary to know about the entities of *Close but Not Frequent* category. The entities belonging to *Close but Not Frequent* category provides in-depth information about the event from the grass root level. The *Close but Not Frequent* entities are likely to be found buried in the Long Tail sources. In order to gain maximum information about an event it is necessary to identify the entities from both the categories. They point to the key persons, places and organizations related to the event. In other words, these entities when present in a source makes it highly specific to the event. Both these categories of entities can prove to be vital sources of information about the event.

7.2 Event-specific and Event-class specific dictionaries

Based on the categorization and the analysis conducted in the previous subsection we divide the final event dictionaries (ς_{E_j}) for each event into *event-specific* and *event-class specific* dictionaries. The *event-specific* dictionaries are comprised of the entities categorized as *Close and Frequent*, and *Close but Not Frequent* for each event $E_j \in \xi$ respectively. Whereas, the *event-class specific*

dictionary is comprised of the entities found in the *Not Close but Frequent* and *Neither Close nor Frequent* categories for all the events $E_j \in \xi$.

Table 2 shows the top five entities in the *event specific* dictionaries for each event E_j and a socio-political event dictionary for the set of events ξ under study. The entities in the *event specific* dictionaries, when present in a source makes it highly specific to that particular event and contributes in gaining information about it. These entities can also help in conducting micro-analysis of an event by identifying precise details about the event. On the other hand, the entities of the *event-class specific* dictionary provides shallow information about a specific event, and are useful in learning about a category of events, in this case, socio-political uprisings in the middle east. These entities can help in conducting macro-analysis of an event by classifying the event into a certain category (like socio-political, crisis, entertainment, economic, etc.) and point out the general characteristics of the event. However, the analysis requires a set of known events, which could be a perceiveable constraint. We plan to address this in the future work.

8 Conclusions & Future Work

In this chapter, we highlighted the need for exploring the social media sources to study an event. We demonstrated that social media sources that are often buried in the Long Tail have the capability to provide very specific and novel information. However, the sheer volume of social media sources and the Long Tail characteristics (e.g., link sparsity, colloquial language, etc.) make it extremely challenging to identify the specific sources. Towards this direction, we developed a methodology that utilizes relevant entities as a mechanism to identify specific sources using a mutual reinforcement framework. Further, in order to consider the dynamic relationship between the specific sources and close entities, an evolutionary mutual reinforcement model is developed. Experiments conducted on real-world datasets for the three social movements during the Arab Spring, viz., Egyptian Revolution, Libyan Revolution, and Tunisian Revolution demonstrate faster convergence and better accuracy of the evolutionary mutual reinforcement model over the conventional mutual reinforcement model. Furthermore, the evolutionary mutual reinforcement model outperformed one of the most-widely used search engines, i.e., Google Blog Search and IceRocket. It was observed that the search engines ranked the specific sources surprisingly low, thereby reducing the chances of their discovery. The poor hyperlink connectivity of these Long Tail social media sources was contemplated to be a primary reason behind their low ranks in traditional search engines. By analyzing the close entities identified by our model we also showed the potential of the framework to be utilized for analyzing events. As next steps, we plan to systematically study the applicability of the model in other social media platforms, especially the microblogging sites (e.g., Twitter).

Recently, social media sources are increasingly found to be plagued with false information that poses challenges for information and knowledge extraction for

an event. The instances of false information are categorized as misinformation (i.e., inadvertent dissemination of false information due to relying upon inadequately validated facts, evolving nature of details during breaking events, or lack of exercising discretion before sharing information, etc.) or disinformation (i.e., intentional falsification of the information or distortion of facts due to disinformation campaigns, aiming to create deception, promoting an ideology or agenda, or false propaganda, etc.). This presents an opportunity for future explorations to examine and further develop the proposed methodology in assessing credibility of social media sources and isolate the ones that are more likely to disseminate false reports.

Acknowledgments

This research is funded in part by the National Science Foundation's Social Computational Systems (SoCS) and Human-Centered Computing (HCC) research programs within the Directorate for Computer & Information Science & Engineering's (CISE) Division of Information & Intelligent Systems (IIS) (Award Numbers: IIS-1110868 and IIS-1110649) and the US Office of Naval Research (Grant numbers: N000141010091 and N000141410489). We would like to thank the Advances in Social Network Analysis and Mining (ASONAM) 2013 conference chairs for inviting us to develop our research further and submit the research to this publication. We are also grateful to the anonymous reviewers for their invaluable comments.

References

1. L. Adamic et al. Power-law distribution of the world wide web. *Science*, 287(5461):2115–2115, 2000.
2. N. Agarwal, M. Lim, and R. T. Wigand. Finding her master's voice: the power of collective action among female muslim bloggers. In *ECIS*, 2011.
3. N. Agarwal, M. Lim, and R. T. Wigand. Online collective action and the role of social media in mobilizing opinions: A case study on women's right-to-drive campaigns in saudi arabia. In *Web 2.0 Technologies and Democratic Governance*, pages 99–123. Springer, 2012.
4. N. Agarwal, M. Lim, and R. T. Wigand. Raising and rising voices in social media. *Business & Information Systems Engineering*, 4(3):113–126, 2012.
5. N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *Proceedings of the international conference on Web search and data mining (WSDM)*, pages 207–218. ACM, 2008.
6. C. Anderson. *Long Tail, The, Revised and Updated Edition: Why the Future of Business is Selling Less of More*. Hyperion, 2008.
7. H. Becker, M. Naaman, and L. Gravano. Selecting quality twitter content for events. *ICWSM*, 11, 2011.
8. J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *Proceedings of the 18th international conference on World wide web*, pages 51–60. ACM, 2009.

9. S. Brin et al. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.
10. N. Diakopoulos et al. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 2451–2460. ACM, 2012.
11. B. Ekdale et al. From expression to influence: Understanding the change in blogger motivations over the blogspan. *AEJMC, Washington, DC*, 2007.
12. G. Erkan et al. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR)*, 22:457–479, 2004.
13. G. Golub et al. *Matrix computations*, volume 3. Johns Hopkins University Press, 1996.
14. M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter. In *SDM*, pages 153–164. SIAM, 2012.
15. N. Hamdy et al. Framing the egyptian uprising in arabic language newspapers and social media. *Journal of Communication*, 2012.
16. Z. Harb. Arab revolutions and the social media effect. *M/C Journal*, 14(2), 2011.
17. T. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):784–796, 2003.
18. A. Jadhav, H. Purohit, P. Kapanipathi, P. Ananthram, A. Ranabahu, V. Nguyen, P. N. Mendes, A. G. Smith, M. Cooney, and A. Sheth. Twitris 2.0: Semantically empowered system for understanding perceptions from social data. *Semantic Web Challenge*, 2010.
19. T. Johnson et al. Wag the blog: How reliance on traditional media and the internet influence credibility perceptions of weblogs among blog users. *Journalism & Mass Communication Quarterly*, 81(3):622–642, 2004.
20. P. Jurczyk and E. Agichtein. Hits on question answer portals: exploration of link analysis for author ranking. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 845–846. ACM, 2007.
21. J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.
22. S. Kumar, G. Barbier, M. A. Abbasi, and H. Liu. Tweettracker: An analysis tool for humanitarian and disaster relief. In *ICWSM*, 2011.
23. A. Langville et al. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380, 2004.
24. L. Liu, L. Sun, Y. Rui, Y. Shi, and S. Yang. Web video topic discovery and tracking via bipartite graph reinforcement model. In *Proceedings of the 17th international conference on World Wide Web*, pages 1009–1018. ACM, 2008.
25. LOmariba. Is new media posing a serious challenge to traditional media?". Technical report, University of Westminster, 2009.
26. D. Mahata and N. Agarwal. What does everybody know? identifying event-specific sources from social media. In *CASoN*, pages 63–68, 2012.
27. D. Mahata and N. Agarwal. Learning from the crowd: an evolutionary mutual reinforcement model for analyzing events. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 474–478. ACM, 2013.
28. A. Marcus et al. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 227–236. ACM, 2011.

29. F. Morstatter, S. Kumar, H. Liu, and R. Maciejewski. Understanding twitter data with tweetxplorer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1482–1485. ACM, 2013.
30. J. Ramos. Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, 2003.
31. T. Rattenbury et al. Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110. ACM, 2007.
32. S. Reese et al. Mapping the blogosphere professional and citizen-based media in the global news arena. *Journalism*, 8(3):235–261, 2007.
33. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.
34. J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51. ACM, 2009.
35. V. Singh et al. Mining the blogosphere from a socio-political perspective. In *Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on*, pages 365–370. IEEE, 2010.
36. R. Troncy et al. Linking events with media. In *Proceedings of the 6th International Conference on Semantic Systems*, page 42. ACM, 2010.
37. A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.