Lecture Notes in Social Networks

Jalal Kawash *Editor*

Online Social Media Analysis and Visualization

Springer

Lecture Notes in Social Networks

More information about this series at http://www.springer.com/series/8768

Jalal Kawash
Editor

Online Social Media Analysis and Visualization

Springer

Editor
Jalal Kawash
Department of Computer Science
University of Calgary
Calgary, AB
Canada

ISSN 2190-5428 ISSN 2190-5436 (electronic)
Lecture Notes in Social Networks
ISBN 978-3-319-13589-2 ISBN 978-3-319-13590-8 (eBook)
DOI 10.1007/978-3-319-13590-8

Library of Congress Control Number: 2014956485

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

Online Social Media (OSM) have revolutionized the way people interact and share information. Many recent events and developments have shown that OSM are very powerful tools for people to organize and take action. Examples include the 'Occupy' movement, 'Sandy' relief efforts, and the 'Arab Spring'. OSM have offered a real and viable alternative to conventional mainstream media. The latter is often accused of being biased. Spin and constraints imposed by regulating or funding bodies can hinder mainstream media outlets' unbiased reporting. They often omit (intentionally or otherwise) certain details in their reporting. On the other hand, OSM are likely to provide "raw", unedited information and the details can be overwhelming with the potential of misinformation and disinformation. Yet, OSM are leading to the democratization of knowledge and information. OSM is allowing almost any citizen to become a journalist reporting on specific events of interest. This is resulting in unimaginable amounts of information being shared among huge numbers of OSM participants. As of this writing, twitter claims to have 271 million monthly active users, producing 500 million tweets per day. Facebook grew by the end of 2013 to 1.23 billion users with 757 million users logging on every day. Facebook now has a user base comparable to the population of India! The Daily Mail further reports that the average American daily spends 40 min on Facebook. This is resulting in several billion "likes" and several 100 million posted pictures in a single day.

This book contains 10 contributions that tackle challenges of the subject of OSM analysis and visualization from different angles. These challenges include:

1. Many details are hidden in OSM posts and as a result, search engines tend to favor conventional media sources. Mechanisms that allow for delving deeper into the content of OSM posts so that a more meaningful search can be performed are needed. How can sources of a Twitter event, for example, be accurately identified?
2. Word-of-mouth marketing is a byproduct of OSM. However with OSM, there is the challenge of linking opinions of users to their demographic information. How can we infer such information with the little clues available on a social

network, such as Twitter? Can the gender of users be inferred based on colors used in Twitter profiles?
3. Interests of OSM participants change over time. How can we analyze this change and how can we present it in a convenient manner? What factors are essential for trend-prediction on Twitter, for instance?
4. With the large sizes of users personal networks (an average Facebook user has more than 220 friends), better ways for OSM users are needed in order to better view their social networks. How can the network be visualized in such a way that gives the user immediate important information about friends and posts, such as the level of activity of a friend and the popularity of a post?
5. Making privacy decisions (what to share with whom) is cumbersome given the sizes of personal networks. Going through the list of friends, one-by-one, may no longer be a viable solution. How can we make such privacy decisions aided by a visualization and categorization mechanism offering different privacy levels for different categories?
6. Network analysis is a complex process. This challenge is approached on two fronts: how can we (a) develop realistic models that describe people's interaction and (b) defuse the complexities of the processes through the development of appropriate tool?

Summary of Contributions

Identifying Event-Specific Sources from Social Media

Identifying valuable sources of social media events is a very challenging task, but is a necessary step toward identifying misinformation and disinformation in social media. These sources are often buried in the "long tail." A quick search for some event on major search engines, such as Google, yields top hits for mainstream and conventional media. In addition, conventional media does not often include as much details as social media alternatives. In this chapter, Debanjan Mahata and Nitin Agarwal take on the challenging task of identifying sources from social medial for specific events. The challenges include sparsity of resources, quality assessment, entity extraction, and evaluation measures.

Mahata and Agrawal develop a mutual reinforcement-based methodology for this identification, present an evaluation strategy, validate the approach using real-life data, and conclude with analysis of the model. The empirical evaluation is based on a data set of 11,378 blog posts from different blogging sources. The events are the Egyptian, Libyan, and Tunisian uprisings. Empirical results show that the developed evolutionary mutual reinforcement converges faster with better accuracy than the conventional mutual reinforcement model, and it outperforms Google Blog Search and IceRocket.

Demographic and Psychographic Estimation of Twitter Users Using Social Structures

With the ever expanding number of social media users, such as Twitter and Facebook, many of their posts are geared toward expressing opinions about certain products and services. This can provide a low-cost, real-time word-of-mouth marketing, as opposed to expensive, formal customer surveys. However, formal surveys have the advantage of linking opinions to customer attributes (such as age and gender), but such attributes are often hidden in social media. In this chapter, Jun Ito, Kyosuke Nishida, Takahide Hoshide, Hiroyuki Toda, and Tadasu Uchiyama analyze more than 4.6 million Twitter users in Japan. It is determined that very few users use values in their Twitter profiles that can reveal their age (roughly 3 % described their age), gender (less than 8 % revealed their gender), location (less than 25 % revealed their location), and occupation (less than 14 % indicated their job). Hence, the estimation of these attributes is necessary.

To address this limitation, Ito et al. provide in this chapter a method by which hidden attributes can be estimated from the publicly available information, Twitter profiles and posts, and from social neighbors. Specifically, Ito et al. estimate the four attributes: gender, age, occupation, and interest. The estimation proceeds at three levels. First, a labeling method that identifies users with blog accounts is used to extract their attributes from the blog profile as true labeling for the training data set. Experiments confirm that this is a more accurate labeling than other methods, such as manual labeling and pattern matching. Second, the authors investigate how to combine bag-of-words features of profile documents and tweets. Nine different combining methods are evaluated, identifying the best two of these methods. Finally, information from social neighbors is utilized. Three adjustment levels are investigated.

Say It With Colors: Language-Independent Gender Classification on Twitter

Gender prediction of social network users is important for targeted advertising, law enforcement, and other social reasons. Gender classification in networks such as Twitter heavily depends on analyzing the text of posted messages or tweets. Among the limitations of such an approach is language-dependence, intractability, and non-scalability.

In this chapter, Jalal Alowibdi, Ugo Buy, and Philip Yu present a gender classification approach that is based on colors. This approach is based on analyzing five color features used in Twitter user profiles: background, text, link, sidebar fill, and sidebar boarder colors. Colors are language independent, and using only five features in the analysis (as opposed to millions of features used in text-based classification) makes the color-based approach more desirable for scalability and

tractability. Realizing that the number of colors can be technically enormous, while practically there can be different grades of the same color, Alowibdi et al. empoly a preprocessing step that converts the colors from their RGB (Red, Green, Blue) representation to HSV (Hue, Saturation, Value) representation. The colors are then sorted by their hue and value attributes, providing similar labellings for colors, which are then converted back to their RGB values. This preprocessing step helps improve accuracy and reduce the size of the data set. Empirical results show that the classification accuracy is roughly between 70 and 74 % for different data sets, which is a clear improvement over the 50 % norm. These results are obtained from a data set of about 170,000 users where it was possible to independently verify their genders.

TUCAN: Twitter User Centric ANalyzer

This chapter by Luigi Grimaudo, Han Hee Son, Mario Baldi, Marco Mellia, and Maurizio Munafò takes a text-mining approach to analyzing Twitter posts, using a framework called TUCAN. TUCAN analyzes the tweets of a single, target user over a specific period of time, identifying the interests of that user during the given time window. TUCAN also offers the ability to do a comparison between several users, inferring any common interests. The results are depicted graphically in an intuitive visual representation.

The steps taken in this framework start by projecting a target user's tweets to a time window that Grimaudo et al. call a "bird song". Next, bird songs are filtered and cleaned using known methods in order to eliminate noise and derive general concept terms for the words in the songs. Terms are then scored to identify the important terms for a target user in a bird song. Finally, similarity scores are used to compare two bird songs. The results are provided visually, using colored matrices, where colors distinguish similarity scores.

The authors validate TUCAN by providing an empirical study that includes 740 Twitter users, including 28 public figures, over a period that exceeds two months. This results in analyzing more than 800,000 tweets. Grimaudo et al. also perform a parameter-sensitivity analysis of TUCAN.

Evaluating Important Factors and Effective Models for Twitter Trend Prediction

In this chapter, Peng Zhang, Xufei Wang, and Baoxin answer two important questions related to trend prediction in Twitter. The first question is what content and context factors (or a combination of them) are more important to Twitter trend prediction. The second question is which (if any) is more appropriate for prediction.

To answer these questions, Zhang et al. performed an empirical study using 16.8 million tweets by about 670,000 users over a period of several months. They conducted relevance analysis using tweet content, network topology, and user behavior, addressing the first question. To address the second question, they also performed a prediction performance study for several known prediction models. The analysis concluded that trend factors based on user behavior are more effective in predicting trends, and that the nonlinear state-space models are more suited for prediction.

Rings: A Visualization Mechanism to Enhance the User Awareness on Social Networks

The visualization of someone's social network activities on Facebook is the subject of this chapter by Shi Shi, Thomas Largillier, and Julita Vassileva. The authors take an approach to represent such activities using rings, where the colors and sizes of these rings represent different levels of activities. The result is a tool called *Rings*. Rings allows a user visually and interactively (1) see basic post information, such as the posters' identification and the time; (2) review the activity level of a user; and (3) assess the popularity of posts.

Shi et al. validate Rings using two user studies. The first study assesses if indeed Rings increases user awareness, whether it is useful to users, and how usable the user interface is. The empirical data show general positive results. The second user study delves deeper into Rings validating more specific properties of the system, such as performance, colors, reliability, and other factors.

Friends and Circle—A Design Study for Contact Management in Egocentric Online Social Networks

Due to the large amount of information that a social network user must deal with, managing privacy decisions related to which posts to share with which users becomes an overwhelming process. Users often indicate that they regret certain social network postings. In addition, an average user can easily have more than one hundred friends or followers; for example, a Facebook user had an average of 229 friends in 2013. Going through someone's network, friend-by-friend, in order to decide what level of privacy is required for each friend is not practical. Instead, users tend to categorize their networks allowing a different privacy-level for each category.

Bo Gao and Bettina Berendt target this issue in this chapter by developing an online application, called *FreeBu*, which visualizes a user's network and categorizes his/her friends. Gao and Berendt accomplish this task through several steps.

They present *CircleTree*, a visualization tool that incorporates modularity-based community detection (MOD). A user study is then performed to compare hierarchical MOD with Facebook smart lists. The findings show that hierarchical MOD provides more support for visibility decisions. They also show that graph-based algorithms for community detection are more appropriate than attribute-based algorithms. The authors, then, empirically compare MOD with Generative Model for Friendships (GMF). The study involves ego-networks as follows: 10 from Facebook, 909 from Twitter, and 129 from Google+. Using this data set, it is shown that MOD outperforms GMF. Finally, Gao and Berendt enrich CircleTree with three additional visual interactive methods culminating in FreeBu, exploiting their empirical findings.

Genetically Optimized Realistic Social Network Topology Inspired by Facebook

There is an ever-increasing need to better understand the topological properties of social networks. This requires the development of abstract and generic, and at the same time, flexible and realistic models that describe how people socially interconnect. The synthetic generation of such a topology is very handy to researchers since it provides them with a mechanism to generate social network data with certain specified properties on demand. In this chapter, Alexandru Topirceanu, Mihai Udrescu, and Mircea Vladutiu address this problem of synthetically generating realistic social network topologies. They propose the Genetic-Optimized Social Network (*Genosian*) method. The aim of Genosian is to create accurate replica of friendship models collected from Facebook.

The first empirical observation made by the authors is that realistic Facebook networks share common metrics, in spite of the fact that the networks are diverse in shape and size. It is found that topological metrics of these realistic Facebook networks fall within narrow thresholds. These metrics include: network size, average path length, clustering coefficient, average degree, diameter, density, and modularity. In addition, distribution of degrees, betweenness, closeness, and centrality are looked at.

Genosian uses a Genetic Algorithm (GA) to generate the social network. It starts with a random creation of a collection of communities, inspired by the Watts-Strogatz algorithm. Then, these communities are linked together. The GA optimizes these intra-community edges until the centrality measure of the graph is finessed, aiming at a comparable value to that of real-life Facebook examples. The rewiring of intra-community edges is carried out using GA's natural selection. The empirical results show that on average Genosian produces 63 % more accuracy than the best previous known method.

A Workbench for Visual Design of Executable and Re-usable Network Analysis Workflows

Network analysis is in general a complex process that consists of several steps. Providing social network researchers with tools to assist them in the analysis processes defuses some of these complexities. Tilman Göhnert, Andreas Harrer, Tobias Hecking, and H. Ulrich Hoppe provide in this chapter a social network analysis tool, called *Analytic Workbench*, which offers several advantages over similar tools. The motivation for the Analytic Workbench is rooted in ease of accessibility, support for complex analysis processes in an integrated environment, and explicit representation of analysis workflows. Another motivating factor is the support and ease of integration of additional analysis functions.

The result is a Web-based interface that provides visual representation of multi-step analysis workflows. Explicit representation of analysis workflows yields the ability to reuse workflows, allowing comparative studies with the same analytic methodology. Furthermore, the tool easily provides adaptation of parts of a workflow in another, allowing a researcher to experiment with different algorithms and analytic steps.

Göhnert et al. showcase the Analytic workbench by using blockmodeling analysis on multi-relational networks. The authors also report on a user evaluation study of the tool.

On the Usage of Network Visualization for Multiagent System Verification

In this chapter, Fatemeh H. Fard and Behrouz H. Far take advantage of visualization techniques in order to build an approach for the verification of Multi-Agent Systems (MAS). They make use of social network analysis in order to model the interaction of agents. The purpose of this study is to detect emergent behavior in these networks. An emergent event is an unexpected run-time behavior of the system unforeseen by system designers.

Fard and Far avoid using model checking techniques for detecting emergent behavior due to the scalability drawback of these approaches. Instead with the use of interaction matrices, three networks are derived. The first is a component-level network that describes the agent's behavior. The other two are system-level networks, defining the interaction of agents. The authors illustrate this approach and compare it to other existing approaches through two examples.

Calgary, Canada Jalal Kawash

Contents

Contributors

Nitin Agarwal Department of Information Science, University of Arkansas at Little Rock, Little Rock, USA

Jalal S. Alowibdi Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA; Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

Mario Baldi Narus Inc., Sunnyvale, CA, USA

Bettina Berendt Department of Computer Science, KU Leuven, Heverlee, Belgium

Ugo A. Buy Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA

Behrouz H. Far Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB, Canada

Bo Gao Department of Computer Science, KU Leuven, Heverlee, Belgium

Tilman Göhnert University Duisburg-Essen, Duisburg, Germany

Luigi Grimaudo Politecnico di Torino, Torino, Italy

Andreas Harrer Clausthal Technical University, Clausthal-Zellerfeld, Germany

Tobias Hecking University Duisburg-Essen, Duisburg, Germany

Fatemeh Hendijani Fard Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB, Canada

H. Ulrich Hoppe University Duisburg-Essen, Duisburg, Germany

Takahide Hoshide NTT Service Evolution Laboratories, NTT Corporation, Yokosuka-shi, Kanagawa, Japan

Jun Ito NTT Service Evolution Laboratories, NTT Corporation, Yokosuka-shi, Kanagawa, Japan

Thomas Largillier GREYC Université de Caen - Basse Normandie, Caen, France

Baoxin Li Computer Science and Engineering, Arizona State University, Phoenix, USA

Debanjan Mahata Department of Information Science, University of Arkansas at Little Rock, Little Rock, USA

Marco Mellia Politecnico di Torino, Torino, Italy

Maurizio Munafò Politecnico di Torino, Torino, Italy

Kyosuke Nishida NTT Resonant Inc, Tokyo, Japan

Shi Shi MADMUC Lab, University of Saskatchewan, Saskatoon, Canada

Han Hee Song Narus Inc., Sunnyvale, CA, USA

Hiroyuki Toda NTT Service Evolution Laboratories, NTT Corporation, Yokosuka-shi, Kanagawa, Japan

Alexandru Topirceanu Department of Computers and Information Technology, Politehnica University Timisoara, Timisoara, Romania

Tadasu Uchiyama NTT Service Evolution Laboratories, NTT Corporation, Yokosuka-shi, Kanagawa, Japan

Mihai Udrescu Department of Computers and Information Technology, Politehnica University Timisoara, Timisoara, Romania

Julita Vassileva MADMUC Lab, University of Saskatchewan, Saskatoon, Canada

Mircea Vladutiu Department of Computers and Information Technology, Politehnica University Timisoara, Timisoara, Romania

Xufei Wang Computer Science and Engineering, Arizona State University, Phoenix, USA

Philip S. Yu Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA

Peng Zhang Computer Science and Engineering, Arizona State University, Phoenix, USA

Identifying Event-Specific Sources from Social Media

Debanjan Mahata and Nitin Agarwal

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

D. Mahata · N. Agarwal (✉)
Department of Information Science, University of Arkansas at Little Rock, Little Rock, USA
e-mail: nxagarwal@ualr.edu

D. Mahata
e-mail: dxmahata@ualr.edu

J. Kawash (ed.), *Online Social Media Analysis and Visualization*,
Lecture Notes in Social Networks, DOI 10.1007/978-3-319-13590-8_1

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.

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

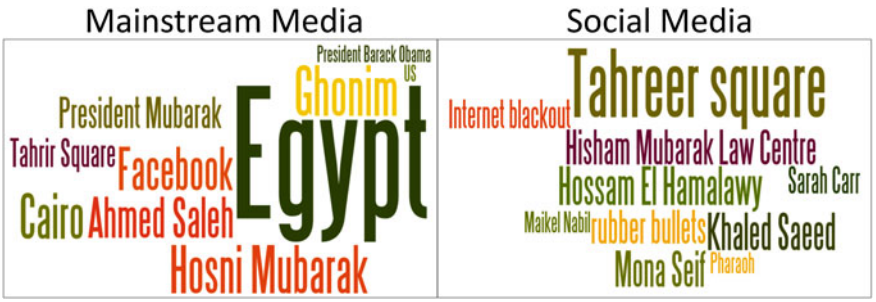

Fig. 1 Top 10 entities from mainstream media and blogs

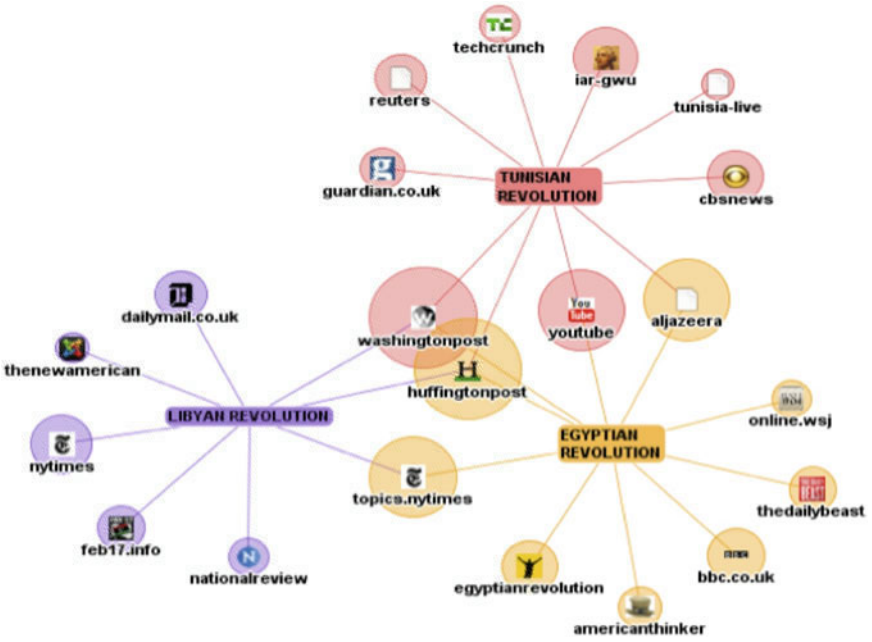

Fig. 2 Top 10 Google search results for "Egyptian Revolution", "Libyan Revolution", and "Tunisian Revolution", visualized using TouchGraph

'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.

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.

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

[1] http://touchgraph.com.

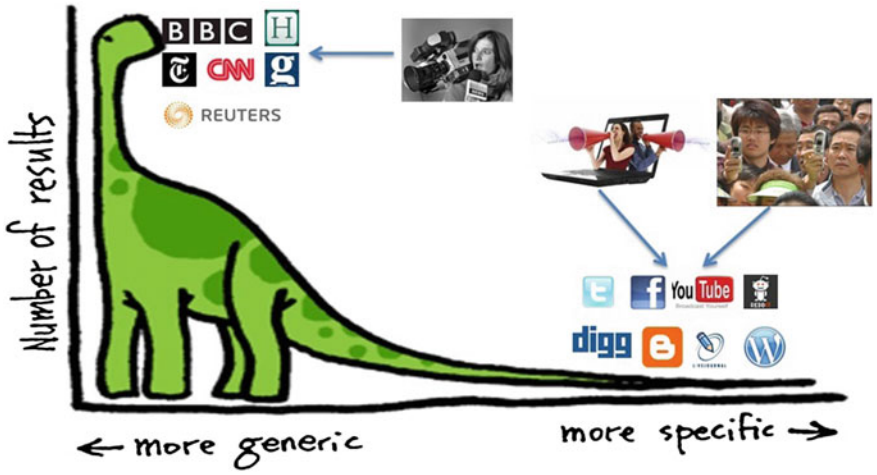

Fig. 3 Short Head versus Long Tail media sources

- **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.

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 (Sect. 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 (Sect. 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' (Sect. 6). However, the work is extendible to other types of events.
- **Event analysis**. Explore the utility of such a model in analyzing events (Sect. 7) and conclude the work with future directions (Sect. 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 Sect. 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 Sect. 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 visualyyizations 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 Sect. 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:

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

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

where, $S_i (\in \phi_{E_j})$, *is the* ith *source, and* $e_i (\in \sigma_{E_j})$ *is the* ith *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:*

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

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

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.

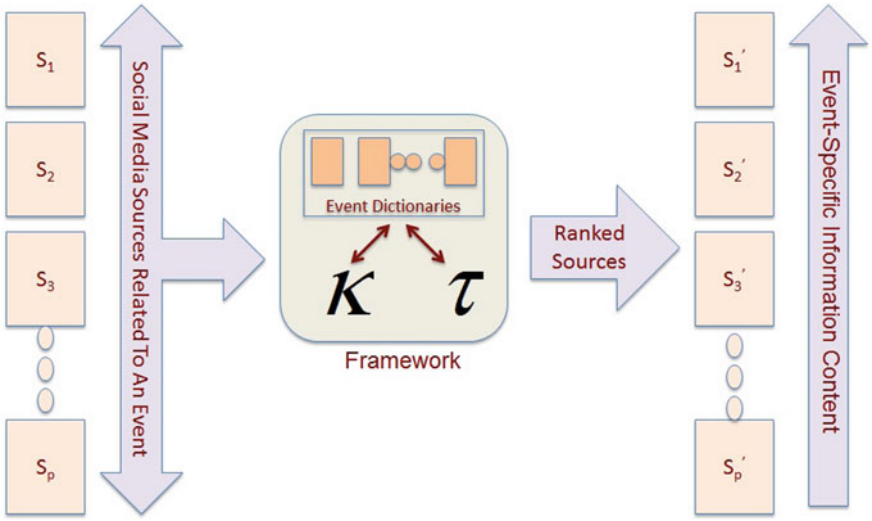

Fig. 4 Black box view of the problem

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 [26, 27].

> 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.

Fig. 5 Entities associated with a social media source

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 Eqs. 5 and 6 based on tf-idf measure [30] , from the information retrieval literature. Let $E_j \in \xi$, be the jth event, and 'e_i' be the ith 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,

$$\tau(e_i)_{E_j} = e_if\_IE_jf = f(e_i, E_j) * IE_jf(e_i) \tag{5}$$

$$IE_jf(e_i) = \log\left(\frac{|\ \xi\ |}{|\ E_j \in \xi : e_i \in E_j\ |}\right) \tag{6}$$

and, $| E_j \in \xi : e_i \in E_j |$ 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 $| E_j \in \xi : e_i \in E_j | > 0$.

We get $| \xi |$ 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 Sect. 5, using AlchemyAPI[3] and their corresponding $\tau(e_i)_{E_j}$ values are calculated using Eq. 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 Eqs. 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.
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.

[2] http://globalvoicesonline.org.

[3] http://alchemyapi.com.

[4] http://docs.python.org/2/library/difflib.html.