A Tool for Assisting Provenance Search in Social Media

Suhas Ranganath, Pritam Gundecha, and Huan Liu
Arizona State University, Tempe, AZ, USA
{Suhas.Ranganath, Pritam.Gundecha, Huan.Liu}@asu.edu

ABSTRACT

In recent years, social media sites are witnessing an information explosion. Determining the reliability of such a large amount of information is a major area of research. Information provenance (aka, sources or origin) provides a way to measure the reliability of information in social networks. The main challenge in seeking provenance is the availability of suitable data consisting of sufficient unique propagation paths. Knowledge of the actual propagation paths for a piece of information will be a valuable asset in provenance search. This paper presents a tool for capturing the propagation network of a given tweet or URL (Uniform Resource Locator) in the Twitter network. Researchers can use this tool to collect information propagation data, design effective strategies for determining the provenance, and gain information about the tweet such as impact, growth rate and users influencing the spread. Two case studies are presented to demonstrate the effectiveness of the system for seeking provenance information.

Categories and Subject Descriptors

H.4.0 [**Information Systems Applications**]: General; J.2 [**Social and Behavioral Sciences**]

Keywords

Provenance, Information Propagation, Social Media

1. INTRODUCTION

Social media provides people with an easy to use and inexpensive channel to publish, receive and understand information. Content generation, which was the privilege of a few influential and knowledgeable sources, is now carried out by millions of social media users.Although this facilitates fast transfer of knowledge and sharing of ideas, information reliability remains an important concern. The origin, motivation and latent purposes for a piece of information in social media are not always clear, leading to questions on its reliability. Provenance of a given piece of data, such as origin, and chain of custody and the attributes of the people along the path assist the user in judging the reliability of the information [3].

Provenance is defined as the history and ownership of a valued object. Although mechanisms to seek provenance data has been developed in different fields like databases and the semantic web [6], provenance data for social media information is a less explored area. In social media, provenance mainly relates to path of information flow, referred to as Provenance Path and the attributes of users along the path, referred to as Provenance Attributes [1]. Provenance Attributes relates to the qualification and affiliation of the users along the path, giving an idea about their relevance and reliability for a given piece of information. A tool to seek provenance attribute information is presented in [4]. Provenance Path relates to the origins, custody chain and the ownership of the information, enabling the user to assess the reliability of the source, the relation between the origin and ownership and the users influencing information propagation.

Seeking the provenance path of a given piece of information in social media is a challenging topic of research [1]. Current research for seeking provenance paths uses synthetically generated pathways from network propagation models [3][8][7][5][2]. Actual data of the propagation paths will benefit in seeking provenance in social media.

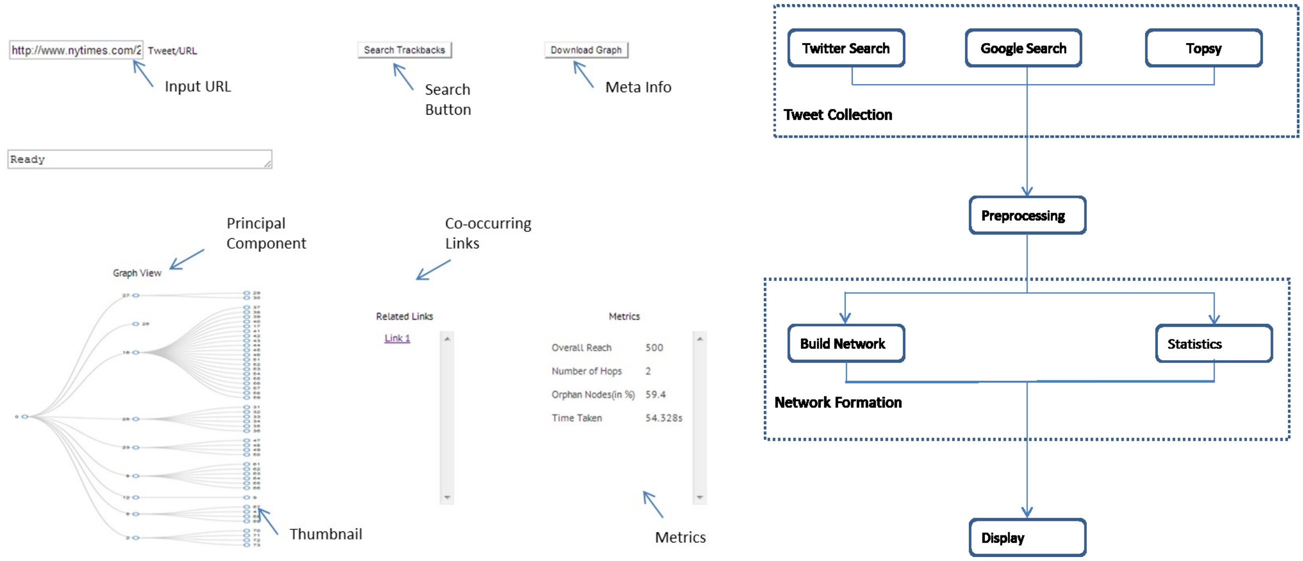
This paper presents a tool for obtaining a spread of given tweet or URL in the Twitter network[1]. In addition to the propagation paths, information about the tweet such as the total observed impact, the number of hops and the most influential users can also be obtained from the framework. The principal component of the propagation network provides an overview of the spread of the tweet or URL and allows the researchers to make additional inferences relevant to their problem. Links co-occurring with the given tweet or URL are also presented whose propagation paths can be accessed on click. This helps in obtaining the propagation path of URLs similar to the given piece of information. The rest of the paper is organized as follows. An overview of the user interface and the system architecture is provided in Section 2. The application of the system in seeking provenance information is demonstrated in Section 3. Concluding remarks with future work is presented in Section 4.

2. PROVENANCE PATH TOOL

This system aims to display the spread of a given tweet or a link across the Twitter network. The tweets and the corresponding users related to the parent tweet or URL's are collected and the users are linked using mentions and retweets. The network is visualized along with basic statistics related to the spread. This framework can be used to understand the popularity, the reach and diffusion patterns of a given tweet. The user interface of the system is described

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.2508214.

[1] The system for detecting Provenance Paths is available online at http://blogtrackers.fulton.asu.edu/Prov_Path/. The demonstration video is available at http://www.screencast.com/t/R9i9xShfigyi.



Overall Reach 500
Number of Hops 2
Orphan Nodes(in %) 59.4
Time Taken 54.328s

(a) User Interface of the system

(b) Block Diagram

Figure 1: System User Interface and Block Diagram

in Section 2.1. A detailed description of the system architecture is described in Section 2.2.

2.1 User Interface

The user interface of the system is illustrated in 1(a). The tweet or URL can be entered in the given search box. The application searches different sources for the given tweets and constructs their propagation network. The principal component of the network can be accessed by clicking on the image below the toolbar. Links which have been mentioned along with the given tweet are also obtained and presented next to the image. The spread of the tweet or link can be obtained by clicking on the hyperlink. The panel on the right hand side presents several useful statistics, such as the spread depth, total coverage and the number of orphan nodes. The "Download Graph" button allows the user to seek provenance information such as edge list, the tweets along the path and the corresponding users. The veracity of the tweets can be determined by using the content of the tweets along the path, while the reliability and neutrality of the users can be visiting the profiles of the users. These data is a valuable asset for determining the reliability of the received information.

2.2 System Architecture

The block diagram of the system is illustrated in 1(b).The system is divided into the following parts: (a) Tweet Collection (b) Data Preprocessing (c) Network Formation (d) Visualization. Each of the modules is described below.

a) Tweet Collection: The tweet or URL is input by the user and related tweets are obtained from various sources on the web. Three different sources are utilized; Twitter search, Google custom search and Topsy. Twitter provides a search service which provides tweets from the past week matching a given query. Google provides a custom search engine service which customizes its search to specific websites. The search was conducted by customizing Google search to twitter.com. Topsy is a social media aggregator which archives tweets and makes them publicly accessible. The main advantage of using Topsy is that older tweets not available from Twitter or Google search can be accessed. Tweets collected from these different sources are aggregated after removing duplicated tweets and are sent to the pre-processing module.

(b) Preprocessing: Two preprocessing steps are implemented. In the first step, symbols such as # and RT and mentioned user ids' are removed. In the next step, similar tweets published by a given user are combined. Comparison of two tweets is done by the following technique. The two tweets are compressed and the compressed size is noted. The two tweets are then combined and compressed and the size of the compressed outputs is compared. The tweets are grouped together if the size of the combined tweet and the two tweets are the same. Nodes in the pathway are then formed as a combination of the tweets and the related username.

(c) Network Module: The network module has 2 components: the network builder and the statistics module. The time ordered list of nodes and tweets from the tweet collection module is used to trace the propagation network of the tweet. The three types of edges are mentions, retweets, and links. The links are formed between the original tweet or URL and the first tweeters in each group. The statistics module presents measures evaluating the spread of the tweet as well as the effectiveness of the system in capturing the spread of the tweet. Some of the measures evaluating the spread of the tweet are tweet popularity given by the total number of people talking about it, propagation depth given by the maximum number of hops in the network the tweet has traveled, and the growth of the spread given by the ratio of the number of users in successive hops. The effectiveness of the system in capturing the tweet is given by the fraction of orphan nodes i.e. nodes which have no edges to the total number of nodes. The output from the network formation module is sent to the display module for network visualization and display of statistics.

(d) Visualization Module:This module displays the spread of the tweet through interactive visualizations. D3, a graph visualization library based on Javascript, is used to visualize the network. Fig 2 shows the principal component of the spread of network for two URLs. The spread of the given piece of information is visualized in the form of a tree, with the root node being the original article. The nodes having children are colored and clicking on the node will display the respective children.

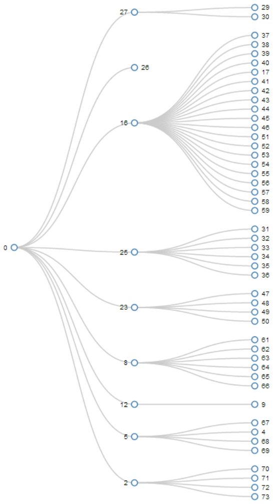


Figure 2: Youtube video of Indonesian riots

3. DEMONSTRATION

The application of the system is demonstrated with two examples. The first example is a link to a video which was used to spread disinformation on Twitter. The second video is on an article reporting the foray of Microsoft in primary education. This tweet was very popular and attracted a lot of traffic.

The first case study is based on a video which was spread in the time of religious riots in North-East India. The video was on a violent attack in Indonesia but was wrongly purported as violence taking place in India. Some users tried to spread disinformation on this video and tried to incite reaction. A recipient of this information can use the proposed system to get the propagation path of the given link. The obtained path is shown in 2. The meta information such as the list of tweets and users are also given. Using this information, users can find the most active propagators of the information and their provenance information such as political affiliation and interests using tools proposed in [4]. For example, the account of one of the most active users, 16, leans very strongly towards hardline nationalism.

The second case study is an article on the New York Times about the role Microsoft is playing in primary education. This tweet has generated a lot of traffic and is a good test for the robustness and performance of the application. Our tool retrieved 500 nodes and inferred propagation paths for more than 42% of the tweets in under a minute. The spread of the tweet is given in 3. From the figure we can observe that a large fraction of the users obtained the information directly from the source. This is in contrast to 2 where most of the users were exposed to the information after a single hop. This might be due to the fact that New York Times is a much reliable source than Youtube. The application gives the user a sense of the reliability of the origin and the chain of custody of information.

4. CONCLUSIONS AND FUTURE WORK

The paper presents a tool to obtain the spread of a given tweet or URL on the twitter network. The provenance path gives additional information to assess the reliability of a given piece of data from social media sites.

The case studies developed to demonstrate the application of the system have given us valuable feedback for improving the system as well as directions for future work. Mechanisms to infer the paths between unconnected nodes in the information pathway need to be investigated. Methodologies need to be developed to assess the value of the tweet based on the obtained spread. The provenance path, once obtained from the data presented in the tool, can pave the way to interesting topics of further research such as information reliability, user credibility and the spread of disinformation in social media.

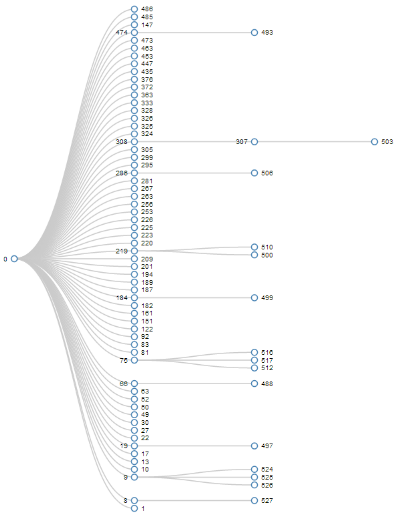


Figure 3: New York Times article on the role of Microsoft in Education

5. ACKNOWLEDGEMENTS

The authors wish to acknowledge the members of the Data Mining and Machine Learning (DMML) laboratory at Arizona State University for constructive feedbacks. This work is sponsored by grants from ARO (#025071) and ONR (N000141110527).

6. REFERENCES

[1] G. Barbier, Z. Feng, P. Gundecha, and H. Liu. *Provenance Data in Social Media. In Synthesis Lectures on Data Mining and Knowledge Discovery*. Morgan & Claypool, 2013.

[2] Z. Feng, P. Gundecha, and H. Liu. Recovering information recipients in social media via provenance. In *The IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2013.

[3] P. Gundecha, Z. Feng, and H. Liu. Seeking provenance of information in social media. In *The 22nd ACM International Conference on Information and Knowledge Management*, 2013.

[4] P. Gundecha, Z. Feng, and H. Liu. A tool for collecting provenance data in social media. In *Demonstration Paper, in the 19th ACM SIGKDD Conference on Knowledge, Discovery and Data Mining*, 2013.

[5] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila. Finding effectors in social networks. In *ACM SIGKDD Conference on Knowledge, Discovery and Data Mining*, 2010.

[6] L. Moreau. The foundations for provenance on the web. *Found. Trends Web Sci.*, 2(3):99–241, Feb. 2010.

[7] B. Prakash, J. Vrekeen, , and C. Faloutsos. Spotting culprits in epidemics: How many and which ones? In *Proceedings of the 12th IEEE International Conference on Data Mining*, 2012.

[8] D. Shah and T. Zaman. Rumors in a network:who's the culprit? *Information Theory, IEEE Transactions on.*, 57(8):5163–5181, 2011.