A Tool for Collecting Provenance Data in Social Media

Pritam Gundecha, Suhas Ranganath, Zhuo Feng, Huan Liu
Computer Science and Engineering
Arizona State University, Tempe, AZ 85281, USA
{Pritam.Gundecha, Suhas.Ranganath, Zhuo.Feng, Huan.Liu}@asu.edu

ABSTRACT

In recent years, social media sites have provided a large amount of information. Recipients of such information need mechanisms to know more about the received information, including the provenance. Previous research has shown that some attributes related to the received information provide additional context, so that a recipient can assess the amount of value, trust, and validity to be placed in the received information. Personal attributes of a user, including name, location, education, ethnicity, gender, and political and religious affiliations, can be found in social media sites. In this paper, we present a novel web-based tool for collecting the attributes of interest associated with a particular social media user related to the received information. This tool provides a way to combine different attributes available at different social media sites into a single user profile. Using different types of Twitter users, we also evaluate the performance of the tool in terms of number of attribute values collected, validity of these values, and total amount of retrieval time.

Categories and Subject Descriptors

H.2.4 [**Information Systems**]: Information Systems Application; J.4 [**Social and Behavioral Sciences**]: Sociology

Keywords

Provenance, Provenance Attributes, Social Media

1. INTRODUCTION

When a social media user receives information via a microblog, a social network, or even a blog site, it is not always clear where the received information originated from, what motivated its publication, and what latent purposes may be associated with the particular information. In such circumstances, a user when presented with additional attribute values can make a better informed judgment about the received information. For example, when the name, occupation, education level, and age can be associated with the originator of information, a user is better informed *about* the received information. In a particular domain, such as politics, a user may be interested in additional attributes. For example, a user with political interests may add political affiliation and special interests to the list of desired attributes.

General Demographic Attribute Set	Domain Specific (Political) Attribute Set
Formal Name (Individual or Group)	Formal Name (Individual or Group)
Location	Location
Occupation	Occupation
Education	Education
Age	Age
	Employer
	Political affiliation
	Religious affiliation
	Lobby affiliation
	Special interest(s)
	Citizenship
	Ethnicity
	Gender

Table 1: Two Lists of Provenance Attributes

In this paper, we present a novel web based tool for collecting attribute values of interest associated with a particular social media user. We refer to these attributes as *Provenance Attributes* and the tool as *Provenance Data Collector*. Currently this tool is designed to assist social media users to collect provenance data of more than half a billion Twitter users[1], and more than a billion Facebook users[2]. Provenance data of a given Twitter or Facebook user is collected from multiple social media sites including Twitter, Facebook, LinkedIn, Wikipedia, and results from Google and Bing search engines.

Specifying the particular set of provenance attributes that are of interest forms the foundation for provenance (sources or originators) search in social media [2]. In practice, sets of provenance attributes are defined subjectively based on the interest of a recipient. As will be shown, some attribute values are easier to obtain than others, and some attribute values may be more important to a recipient than other attribute values. For example, a statement published by a political candidate might be assessed with some bias, if the recipient knows information about the candidate such as party affiliation or special interest associations. Perhaps an even more interesting example of the importance of prove-

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.
Copyright 2013 ACM 978-1-4503-2174-7/13/08 ...$15.00.

[1] http://en.wikipedia.org/wiki/Twitter
[2] http://en.wikipedia.org/wiki/Facebook

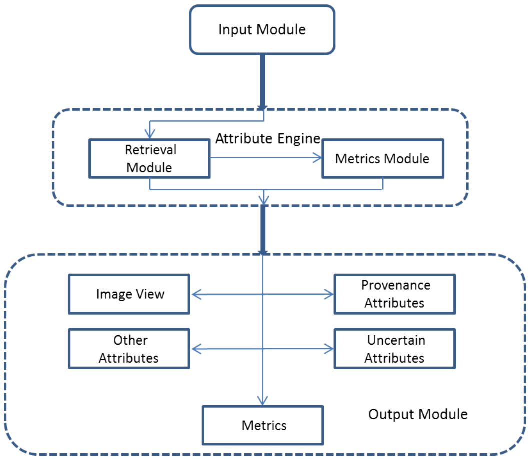


Figure 1: Overview of the Tool for Collecting Provenance Attribute Values.

nance attribute values would be revealing the political affiliation and special interests of an unfamiliar social media user propagating political statements to better judge any latent motivations for propagating a statement in social media. Table 1 displays a "general" and a "domain-specific" attribute list, the provenance attribute sets analyzed in [2]. The attribute sets in Table 1 are grounded in standard demographic information [3].

Often a user has multiple accounts on different social media sites, and is unaware of how much information about her can be publicly available for everybody. For a given user, this tool is an example of how to combine different attributes from different social media sites to form a single user profile. In terms of privacy, vulnerability of a user depends on publicly available personal and sensitive attributes [4]. The proposed tool can also be used to show how vulnerable a user is on social media.

2. PROVENANCE DATA COLLECTOR

Provenance data collector[3] is an online data collection tool focusing on efficiently retrieving useful attribute values of a given Twitter or Facebook user. This tool features an intuitive user interface and is designed to enable fast retrieval of a maximum number of desired provenance attributes. If some desired provenance attributes are uncertain, the tool provides best possible URL (Uniform Resource Locator) to help users further their findings. In addition to provenance attributes, the tool also presents other attribute values and related images during the search, and measures to evaluate efficiency of the system. Figure 2 shows an overview of the tool for collecting provenance attribute values. The following is a detailed description of the tool consisting of three major components.

2.1 Input Module

This module asks social media users to perform two tasks: input the user (Twitter or Facebook) identifier and, select attributes of interest from a list of attributes. We use the Twitter handle and Facebook username to uniquely identify each Twitter and Facebook user, respectively. Each Twitter handle is prefixed by "@". For example, the unique Twitter handle for President Barack Obama is "@barackobama". User's selection of attributes is referred to as *provenance attributes*. The attribute engine then uses the Twitter handle or Facebook username and provenance attributes to retrieve useful information.

2.2 Attribute Engine

The attribute engine is at the core of the provenance data collector tool. The primary objectives are to retrieve values of provenance attributes and compute measures to evaluate efficiency of the tool.

Attribute Retrieval Module

This module takes user identifier and provenance attributes from the input module and explores different social media sites for information collection. The attribute retrieval module utilizes five main important social media sites for mining the attribute values: Twitter or Facebook public profile, LinkedIn public profile, Wikipedia page, and search engine results from Google and Bing. Using the user identifier, formal name and location can be obtained from the Twitter or Facebook user's profile. Both attribute values are then queried on different search engines, including Google and Bing. The user profile from professional social networking sites such as LinkedIn is retrieved from the search results. The publicly available LinkedIn profile page is then mined for the attributes. LinkedIn consists of 200+ million users[4]. We find that, if a Twitter user is also available on LinkedIn, many attribute values can be collected from their publicly available profile page. Some popular users also have their Wikipedia pages which are also used by the module to obtain attribute values. Attributes obtained from the above sites require different information retrieval techniques.

In addition to provenance attributes, we also keep records of other collectible attribute values from the visited sites. To provide further authentication for the collected attribute values, we retrieve related images of a user using results from Google and Bing search engines.

Metrics Module

Based on all the attribute values collected by the retrieval module, the metrics module computes three measures: information availability, information legitimacy, and retrieval speed. These measures help us to evaluate the efficiency of our system as well as provide a way to compare and contrast the information regarding different types of input users.

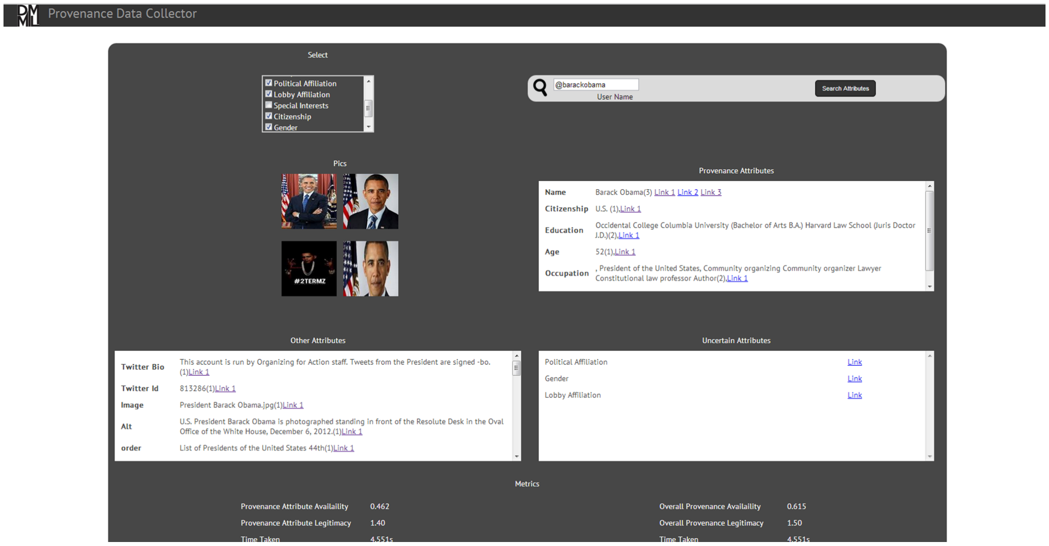
Provenance availability. This metric measures the amount of information available about an input user (α). W is the set of weights, $(w_1, \ldots, w_n) \in W$, associated with provenance attributes $(a_1, \ldots, a_n)$. V_α is the set of provenance attribute values $(v_1, \ldots, v_n) \in V$, associated with α. In order to quantify progress in obtaining the provenance attribute values, an *availability* function, $A : V_\alpha \to [0, 1]$, is defined as

$$A(V_\alpha) = \frac{\sum_{i=1}^{n} w_i \times x_i}{\sum_{i=1}^{n} w_i}, \qquad (1)$$

where $x_i = 0$ if v_i is unknown; otherwise, $x_i = 1$. Weighing particular provenance attributes depends on the domain and the purpose of collection. Currently, the tool gives equal

[3] The provenance data collector tool is located at `http://blogtrackers.fulton.asu.edu/Prov_Attr`, and demonstration video can be found at `http://www.screencast.com/t/XujEYbBXBKBd`

[4] `http://en.wikipedia.org/wiki/LinkedIn`


Provenance Data Collector

Select

☑ Political Affiliation
☑ Lobby Affiliation
☐ Special Interests
☑ Citizenship
☑ Gender

@barackobama
User Name

Provenance Attributes

Name	Barack Obama(3) Link 1 Link 2 Link 3
Citizenship	U.S. (1),Link 1
Education	Occidental College Columbia University (Bachelor of Arts B.A.) Harvard Law School (Juris Doctor J.D.)(2),Link 1
Age	52(1),Link 1
Occupation	, President of the United States, Community organizing Community organizer Lawyer Constitutional law professor Author(2),Link 1

Other Attributes

Twitter Bio	This account is run by Organizing for Action staff. Tweets from the President are signed -bo. (1)Link 1
Twitter Id	813286(1)Link 1
Image	President Barack Obama.jpg(1)Link 1
Alt	U.S. President Barack Obama is photographed standing in front of the Resolute Desk in the Oval Office of the White House, December 6, 2012.(1)Link 1
order	List of Presidents of the United States 44th(1)Link 1

Uncertain Attributes

Political Affiliation	Link
Gender	Link
Lobby Affiliation	Link

Metrics

Provenance Attribute Availaility	0.462	Overall Provenance Availaility	0.615
Provenance Attribute Legitimacy	1.40	Overall Provenance Legitimacy	1.50
Time Taken	4.551s	Time Taken	4.551s

Figure 2: Web Interface of the Provenance Data Collector Tool Showing Provenance Attribute Values of President Barack Obama (@barackobama).

weights to all selected provenance attributes. The availability function describes how many provenance attributes are available for the user of interest. The availability function allows a user to perform simple comparisons of search strategies that are employed to seek provenance attributes. Additionally, the availability function allows a recipient to prioritize search results. Specific user applications designed to obtain provenance attributes can be compared based on the number of attribute values found.

Provenance Legitimacy. Finding provenance attribute values can provide meaningful insights to a social media user, but we also want to know if we have found the valid values of provenance attributes for an input user α. The validity of the attributes can be verified by matching the values from independent social media sites. The legitimacy function is computed by averaging the number of independent social media sites used to verify the attribute. Let I_{V_α} be attribute value *site counters*, $(c_1 \ldots c_n)$, for provenance attribute values in the corresponding V_α. The *legitimacy* function, $l : I_{V_\alpha} \to \mathcal{R}$, is proposed to quantify whether or not the provenance attribute values found are valid.

$$l(I_{V_\alpha}) \quad = \quad \frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n} x_i},$$

where $x_i = 0$ if v_i is unknown; otherwise, $x_i = 1$.

Retrieval time. This metric computes the time taken to obtain information about the given user. It signifies the speed at which information can be retrieved from different social media sites. This is used to evaluate the ease of data retrieval of a particular subject and also the efficiency of the system. In Section 3, we will discuss the performance of the collection tool based on these measures.

2.3 Output Module

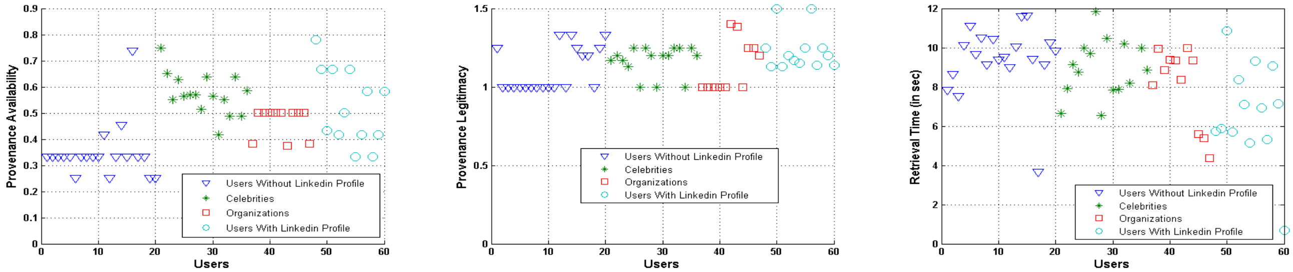
The output module obtains the attribute values and the metrics, segregates them into categories and presents them in easily readable formats. The output module is divided into five sections; four sections corresponding to different attribute categories and one presenting the provenance metrics. In Figure 2, we see the web interface for the provenance data collector tool showing provenance attribute values of President Barack Obama (@barackobama).

The first (Pics) section shows the images related to an input user. Visual information plays a significant role in shaping user confidence about the values of collected provenance attributes. The second (provenance attributes) section displays the values of those provenance attributes which can be found using our retrieval model. The number of social media sites from which the particular value is verified along with the URLs to the site are presented alongside each provenance attribute. The third (other attributes) section shows additional attributes retrieved from different sites. These attributes, although not specifically asked for by the user, present diverse viewpoints about an input user. Provenance attributes with uncertain values are presented in the fourth (uncertain attributes) section. In this case, the user is directed to the most relevant URL where she might be able to find more information. Information available in this section is an indication of information that is hard to retrieve. The last section displays the provenance metrics.

3. SYSTEM EVALUATION

A detailed assessment has been conducted to evaluate the system and to investigate further directions of work. The methodology and the observations, along with the inferences drawn from the evaluations are presented in this section.

An evaluation methodology assessing the proposed tool's various features has been developed. The system is evaluated using measures such as provenance availability, provenance legitimacy, and retrieval time. In order to compare the performance across diverse demographics, different types of Twitter users are input to the system and the evaluation metrics are compared. The pedagogy adopted for the assessments of the proposed tool is as follows: (a) four categories of users, well-known celebrities, normal users with LinkedIn profiles, normal users without LinkedIn profiles, and organizations, are identified, (b) the system gathers the relevant attributes and provenance availability, provenance legitimacy, and retrieval time which are computed for each of the selected Twitter users, and (c) the metrics for the four categories



(a) Availability (b) Legitimacy (c) Retrieval Time

Figure 3: Based on different measures of provenance attributes, performance of the proposed tool for the identified types of Twitter users.

of users are plotted using scatter plots shown in Figure 3. In each scatter plot, we grouped all Twitter users of same type together for simplicity. The computed metrics is then compared to evaluate system performance.

There are in total 60 Twitter users that are selected for system evaluation: 16 celebrities, 13 users with LinkedIn profile, 20 users without LinkedIn profile, and 11 organizations. Provenance availability is computed and compared across different categories of users in Figure 3a. Provenance availability measures the highest for celebrities, which is expected, considering celebrity-related information is easily accessible online. The system is also able to obtain an average of 15 values pertaining to their professional and personal lives, in addition to the identified attributes. A surprising result is that the normal users with LinkedIn profiles rank close to celebrities in provenance availability. These users also rank high in provenance legitimacy, as can be seen in Figure 3b. The system efficiency, as measured by retrieval time, shown in Figure 3c, is approximately equal across all the categories, showing the consistency of the system. Although the system is able to obtain substantial information regarding organizations, provenance availability for organizations is typically low (see Figure 3a). Organizations have varied functions and capabilities; hence, defining homogeneous provenance attributes for organizations is a challenging task.

4. RELATED WORK

Most social media sites, including Twitter, Facebook, and LinkedIn, provide a platform for users to share their attributes. Though users have a choice to keep these attributes hidden, many of these attributes are publicly visible for everyone [4]. Researchers have been using user attributes available in social media sites to address many social media tasks, including deanonymization of networks [7], prediction of private attributes [8], quantification of a user vulnerability [4], and identification of information provenance [1, 5].

Previous research [2] has shown that provenance attribute values could be vital to the task of identifying the provenance of information. They use manual and semi-automatic ways to collect provenance attribute values. In this work, we provide a completely automated way of collecting provenance attribute values from multiple social media sites.

5. CONCLUSIONS AND FUTURE WORK

The provenance data collector tool aims to collect provenance attribute values of a user. By collecting such values of a user related to the received information, the tool could facilitate recipients to understand more about the received information. Data generated on social media sites is largely distributed and unstructured in nature [6]. The proposed tool also provides a way to combine such distributed and unstructured social media data.

The results obtained from the evaluations have provided us with valuable feedback for improving the system as well as directions for future work. Improved methodologies for entity resolution must be investigated to better resolve profiles of a same user across different domains. Better data mining techniques need to be incorporated to resolve uncertain attribute values. Extending the application to search provenance attribute values of a diverse range of entities such as corporations, media, and political parties provides an interesting direction for future work. Designing more robust measures for system and user evaluation is also a compelling area for future investigation.

6. ACKNOWLEDGMENTS

This research is sponsored by the Army Research Office [ARO, #025071].

7. REFERENCES

[1] G. Barbier, Z. Feng, P. Gundecha, and H. Liu. *Provenance Data in Social Media*. Morgan & Claypool Publishers, 2013.

[2] G. P. Barbier. Finding Provenance Data in Social Media. Dissertation, Arizona State University, 2011.

[3] N. M. Bradburn, S. Sudman, and B. Wansink. *Asking Questions*. John Wiley & Sons Inc., 2004.

[4] P. Gundecha, G. Barbier, and H. Liu. Exploiting Vulnerability to Secure User Privacy on a Social Networking Site. In *the 17th ACM SIGKDD*, 2011.

[5] P. Gundecha, Z. Feng, and H. Liu. Recovering Information Recipients in Social Media via Provenance. In *The IEEE/ACM ASONAM*, 2013.

[6] P. Gundecha and H. Liu. Mining Social Media: A Brief Introduction. *Tutorials in Operations Research*, 2012.

[7] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *the 29th IEEE Symposium on Security and Privacy*, 2008.

[8] E. Zheleva and L. Getoor. To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles. In *the 18th ACM WWW*, pages 531–540, 2009.