Data Mining In LinkedIn: Related Research Paper
Danielle Aring & Noreen Halley
A Tool For Collecting Provenance Data in Social Media

Pritam Gundecha, Suhas Ranganath, Zhuo Feng, Huan Liu
Arizona State University, Tempe, AZ 85281, USA
Issues In Social Media Sites: LinkedIn, Facebook, Twitter

❖ Large amounts of user data
➤ This data is both DISTRIBUTED and UNSTRUCTURED!

❖ Unclear information:
➤ Origin
➤ purpose of publication
➤ latent motivations of user statements.

❖ User unable to establish in received information:
➤ trust
➤ value
➤ validity
Solution: PROVENANCE ATTRIBUTES

❖ What Are They?
  ➢ Additional attribute values of interest (provide more depth to user information)
  ➢ **Examples:** political affiliation, lobby affiliation, name, location, age

❖ How do they solve the issues?
  ➢ Aides users to make better informed judgements about an unfamiliar social media user’s motivations
  ➢ Associate attributes with originator of information: user better informed

❖ This paper presents:
  ➢ **Provenance data collector:** Web-based tool to assist users in collection of attribute values of interest
  ➢ Provenance data of each Twitter/Facebook user is collected from LinkedIn, Wikipedia, Bing and Google
Provenance Data Collector

- **Input Module**
  - 2 tasks
    - Identifier (handle/username)
    - Attribute selector

- **Attribute Engine**
  - Core of data collector
    - Retrieve provenance attribute values
    - Compute measures (evaluate tool efficiency)

- **Output Module**
  - Presents single user profile
    - Divided in 5 sections
    - Segregates and presents attribute values/metrics

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

- Measures amount of info available for input user
  - \( W \) is set of weights
  - \( V \) is set of provenance attribute values
  - \( x_i = 0 \) if \( v_i \) is unknown else \( x_i = 1 \)
- Assigns equal weights
- Allows comparison of search strategies
- Prioritizes search results

\[
A(V_\alpha) = \frac{\sum_{i=1}^{n} w_i \times x_i}{\sum_{i=1}^{n} w_i}
\]
Provenance Legitimacy Measure

- Quantifies validity of provenance attributes
- Averages the number of independent social media sites used to verify attribute
  \( l(I_{V_a}) = \frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n} x_i} \)

  ➢ (att site counters) \((c_1 \ldots c_n)\) for values in the same
  ➢ \(X_i = 0\) if unknown attribute
    otherwise \(x_i = 1\)
Retrieval Time

❖ Computes time take to obtain info for given user
❖ Evaluates:
  ➢ Ease of Data retrieval
  ➢ System efficiency
Output Module: Web Interface

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

Test Input:

❖ 4 Categories of Twitter users
   ➢ Celebrities
   ➢ Normal users with Linkedin
   ➢ Normal users without Linkedin
   ➢ Organizations

❖ Input Size - 60 Users
   ➢ Celebrities had highest provenance numbers
   ➢ Normal users with Linkedin very close to Celebrity results
Scatterplots for Test Data Results

(a) Availability

(b) Legitimacy
Scatterplots

(c) Retrieval Time
Provenance Data Collector - Other

- Web-based GUI tool
- Sponsored by the Army Research Office
- Proposed Future Tool Enhancements
  - Improve methods for entity resolution (user profiles across different social media sites)
  - Better data mining techniques to resolve uncertain attributes
  - Extend the application (i.e. Include Corporations, Medial, Political Party searches)
  - Implement more robust measures
  - User evaluation