PROV provenance representing the Twitter data model

Hugo Firth and Paolo Missier

School of Computing Science, Newcastle University, UK
{h.firth, paolo.missier}@ncl.ac.uk

1 Introduction

We describe a large synthetic corpus[1] of PROV data, designed to represent the provenance of artefacts and interactions by users of the popular social network Twitter[2]. The data is topologically valid with respect to the PROV data model [1], and serialised using the PROV-N notation [1]. Table 1 provides key information about the dataset itself.

Table 1: Information about the PROV corpus.

Data format	PROV-N
Data model	PROV
Size on disk	$\sim$ 250 Megabytes
Order |V(G)|	$\sim$ 1.5 Million
Size |E(G)|	$\sim$ 2 Million

We elected to generate the described corpus, rather than collect it from some existing source, for several reasons: *1)* social networking data are not freely available, being of value to their respective service owners; *2)* the scale of social networking data precludes collecting more than a small sample for analysis or benchmarking purposes. This "through the keyhole" approach results in an incomplete dataset, which significantly alters the type of analysis you can run against it.

2 Corpus creation

Our provenance data was created using a modified version of the *provGen*[3] provenance generation tool, with carefully selected parameters designed to create a dataset which exhibited the same statistical characteristics as if it had been collected from Twitter itself. These statistical characteristics of twitter data are determined using an informal report published by a social media analytics firm[4].

[1] https://github.com/provbench/Twitter-PROV
[2] https://twitter.com/
[3] http://prov-gen.com/
[4] http://www.beevolve.com/twitter-statistics/

They are not intended to be rigorous; merely provide the appearance of similarity. The dataset represents the relationships and tweets of 1000 twitter users. The number of followers per user is assigned randomly, according to a Gamma distribution[5]. Figure 1 depicts the anatomies of both the underlying Twitter data model being simulated, and the corresponding PROV representation.

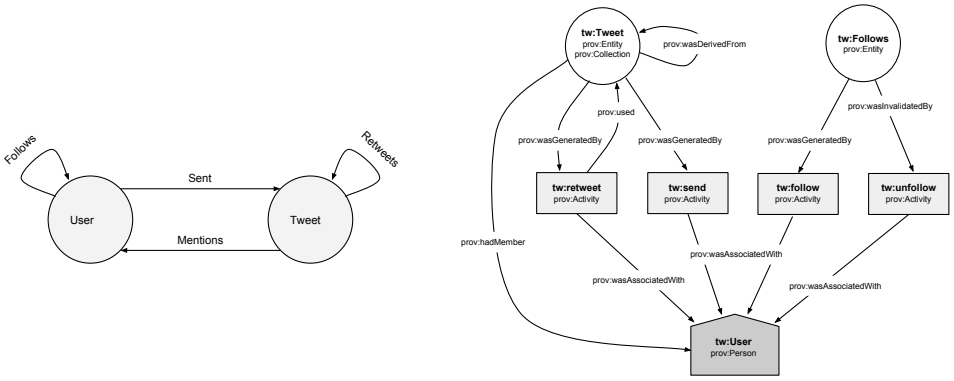

Fig. 1: Twitter data model (left) and its PROV representation.

3 Motivation

Traditional RDBMS, and the applications that make use of them, are primarily interested in "shallow" traversals, such as *"Who are the friends of . . . "* and *"Which entries have x equal to y"*. Provenance data, in part, models the dependencies of data. Retrieving the dependency graph for any one piece of data may require an arbitrarily "deep" traversal. Furthermore, existing RDBMS are often focused on frequent updates of stored data, which may require complex concurrency strategies to ensure the consistency of read operations. PROV data, on the other hand, is monotonic; write operations add new elements, rather than changing existing ones.

We believe that the described dataset can be usefully applied to benchmarking the performance of systems designed to address these PROV specific characteristics, and supply a selection of sample queries to support this.

4 Sample provenance queries

Q1 *What is the longest chain of derivations amongst Tweet Entities (retweets)?* This query might be used to measure the performance of a system at performing a large number of deep traversals of a dependency graph.

[5] http://en.wikipedia.org/wiki/Gamma_distribution

Q2 *What is the most influential Tweet Entity (most widely retweeted tweet)?*
This query might be used to measure the performance of a system at performing broad, shallow, traversals of the type typically found in RDBMS and columnar stores.

Q3 *What is the average number of unfollow Activities associated with a user Agent?*
This query might be used to measure the performance at a system at performing aggregation tasks.

Q4 *What is the great number of on-going Follows Entities attributed to a user Agent at any one time.*
This query might be used to measure the performance of a system in providing answers about the historical state of a dataset.

5 Coverage of PROV

We now address in more detail how nodes and relationships in the Twitter data model map to core PROV concepts.

5.1 PROV Elements

Agent
Agents in the model are the Twitter users who are responsible for changes in the model state. In the described corpus, all activities are associated with an agent with type value *prov:Person*.

Activity
Activities in the model represent actions taken by Twitter users and may take one of the following four forms: *send*, *retweet*, *follow* and *unfollow*. As with Agents this subclassing is achieved via the *prov:type* extensibility point.

Entity
Entities in the model represent Tweets and Follow relationships. The latter has been reified in the PROV representation. Those Tweet entities which mention user agents have the type value *prov:Collection* and have *hadMember* relationships with mentioned agents.

5.2 PROV Relations

used
The retweet activity *uses* a Tweet entity.

wasGeneratedBy
The send activity *generates* a Tweet entity. Further, a follow activity *generates* a Follows entity.

wasDerivedFrom
A Tweet entity, generated by a retweet activity, is *derived* from another Tweet entity.

wasAssociatedWith
All activities are *associated* with user agents.
wasInvalidatedBy
A Follows entity is *invalidated* by an unfollow activity.
hadMember
A Tweet entity which mentions a user agent has that agent as a *member*.

6 Conclusions and further work

In the submission of the described corpus to ProvBench we provide a single provenance dataset for benchmarking purposes. The dataset is generated using specific parameters controlling, for example: the statistical distribution of follower counts; the correlation between variables such as Tweet count and follow count; and the base probability of certain events (such as one user unfollowing another). Given these facts, we can usefully continue to produce further Twitter-PROV datasets. Each generation varying individual parameters in an attempt to quantify the impact of data with different statistical characteristics, upon provenance management systems.

Finally, the use of timestamps in the dataset is restricted to *follow* & *unfollow* activities. This is a result of the complexity involved in calculating correlations between Tweets and Follow relationships over time. However, technical issues not withstanding, this could prove a useful addition to the datasets, and should form the basis for further work.

References

1. Luc Moreau, Paolo Missier, Khalid Belhajjame, Reza B'Far, James Cheney, Sam Coppens, Stephen Cresswell, Yolanda Gil, Paul Groth, Graham Klyne, Timothy Lebo, Jim McCusker, Simon Miles, James Myers, Satya Sahoo, and Curt Tilmes. PROV-DM: The PROV Data Model. Technical report, World Wide Web Consortium, 2012.