CollabMap Provenance: Supporting Quality Assessment and Decision Making

Trung Dong Huynh and Luc Moreau

Electronics and Computer Science, University of Southampton
Southampton, SO17 1BJ, United Kingdom
tdh@ecs.soton.ac.uk, l.moreau@ecs.soton.ac.uk

Abstract. CollabMap is an online crowdsourcing application that was developed to help emergency planners at Hampshire County Council in the UK to create maps for high-fidelity crowd simulations. The main feature of the system is a crowdsourcing mechanism that breaks down the problem of creating evacuation routes into micro-tasks that a contributor to the platform can execute in less than a minute. In this article, we report how provenance of crowdsourced data was recorded by the system and its applications: classifying data quality and supporting decision making.

Keywords: data provenance, crowdsourcing, data quality, mapping, provenance analytics

1 Introduction

CollabMap[1] [6] is a crowd-sourcing platform for constructing evacuation maps for urban areas. These maps need to contain evacuation routes connecting building exits to the road network while avoiding physical obstacles such as walls or fences, which existing maps do not provide. The application crowd-sources the drawing of such evacuation routes from the public by providing them with aerial imagery and ground-level panoramic views. It allows inexperienced users to perform tasks without them needing the expertise to integrate the data into the system. To ensure that individual contributions are correct and complete, the task of identifying routes for a building was broken into different micro-tasks done by different contributors: building identification (outline a building), building verification (vote for the building's validity), route identification (draw an evacuation route), route verification (vote for validity of routes), and completion verification (vote for the completion of the current route set). This allows individual contributors to rate and correct each other's contributions (see [6] for more details).

CollabMap was built in 2011, at that time the PROV standards by W3C [2] have not been published. Instead, we modelled the provenance of crowd-generated data in the application based on the Open Provenance Model (OPM) [5]. An

[1] http://www.collabmap.org/

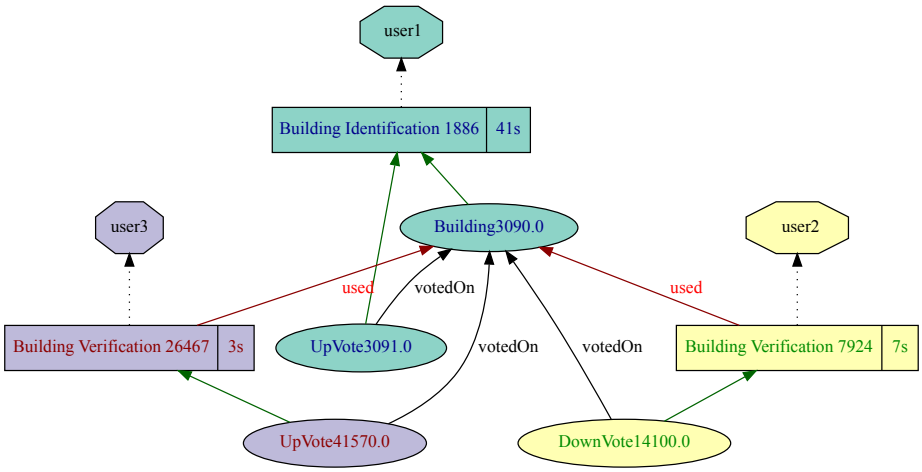

Fig. 1. An example OPM provenance graph recorded by CollabMap showing a building was drawn and voted on by three different users.

example of the provenance is illustrated in Figure 1, which shows a building was identified by `user1` and later verified by `user2` and `user3`, generating votes on the building. Thanks to the significant similarities between the PROV Data Model and OPM, after PROV had been released, we were able to map the recorded provenance to the new PROV Data Model and enabled the application to export its provenance in PROV-JSON [4].

2 Provenance-based data quality classification

During its three-month deployment, over 5,000 buildings were identified with their evacuation routes in CollabMap, the provenance of each building and its routes were recorded in an OPM graph. Depending on the complexity of a building, a provenance graphs in CollabMap can have up to 200 nodes, making it virtually impossible to be analysed visually. In order to address this challenge, network metrics were developed to analyse topological characteristics of such complex provenance graphs [1]. Based on such metrics, machine learning techniques were then employed to explore potential correlations between the topological metrics and properties of data, allowing us to build a predictive model for the quality of data generated by the crowd in CollabMap (see [3] for more details). With this method, we demonstrated that provenance can be used to assess properties of the data it describes with high accuracy (over 95% in the case of data quality in CollabMap).

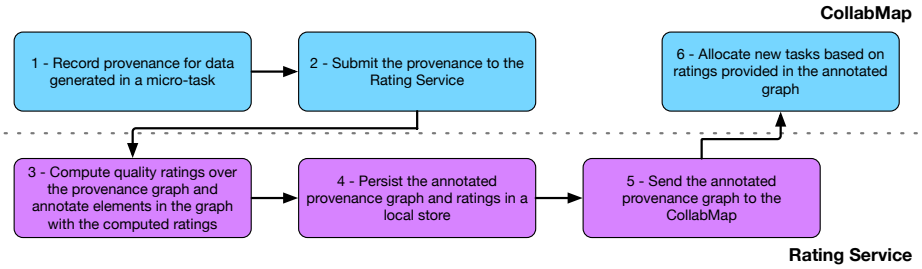

Fig. 2. Overview of the interactions between CollabMap and the Rating Service.

3 Provenance-based online decision-making

In the above approach, provenance graphs are analysed in isolation. As crowd workers in CollabMap typically contributed to several buildings, the data attributed to them also exist in various provenance graphs. In order to assess the reliability of crowd workers, we needed to "connect" their contributions across the graphs. Hence, CollabMap was later extended to delegate the assessment of data quality and user reliability to a rating service (see Fig. 2 for an overview of the process). After each micro-task has finished, the provenance recorded is sent to the rating service (in PROV-JSON). The rating service calculates relevant quality ratings for data entities and crowd workers by propagating votes to data entities and to crowd workers along the provenance graph and aggregating them according to predefined provenance patterns [7]. Elements on the original provenance graphs are then annotated with the computed ratings; the annotated graph is sent back to CollabMap, which makes use of the ratings to decide whether a building is deemed to be finished or more micro-tasks are needed, and to whom those tasks should be distributed. It should be noted that the communications between CollabMap and the rating service are solely in PROV-JSON and the rating propagation and aggregation are carried out based on provenance patterns defined using the PROV Data Model. Therefore, the rating service can be extended or re-purposed to work with a different application following the same approach.

Acknowledgements.

We gratefully acknowledge funding from the UK Research Council for project Orchid, grant EP/I011587/1.

References

1. Ebden, M., Huynh, T.D., Moreau, L., Ramchurn, S., Roberts, S.: Network analysis on provenance graphs from a crowdsourcing application. In: Groth, P., Frew, J.

(eds.) Provenance and Annotation of Data and Processes, Lecture Notes in Computer Science, vol. 7525, pp. 168–182. Springer Berlin Heidelberg (2012)
2. Groth, P., Moreau, L.: PROV-Overview. An Overview of the PROV Family of Documents. W3c working group note, World Wide Web Consortium (Apr 2013), http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
3. Huynh, T.D., Ebden, M., Venanzi, M., Ramchurn, S.D., Roberts, S., Moreau, L.: Interpretation of crowdsourced activities using provenance network analysis. In: First AAAI Conference on Human Computation and Crowdsourcing. pp. 78–85. Palm Springs, CA, USA (2013)
4. Huynh, T.D., Jewell, M.O., Keshavarz, A.S., Michaelides, D.T., Yang, H., Moreau, L.: The PROV-JSON serialization. W3C Member Submission, World Wide Web Consortium (April 2013), https://www.w3.org/Submission/2013/SUBM-prov-json-20130424/
5. Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., Van den Bussche, J.: The Open Provenance Model core specification (v1.1). Future Generation Computer Systems 27(6), 743–756 (jul 2011)
6. Ramchurn, S.D., Huynh, T.D., Venanzi, M., Shi, B.: Collabmap: Crowdsourcing maps for emergency planning. In: 5th ACM Web Science Conference (WebSci '13) (2013)
7. Sezavar Keshavarz, A., Huynh, T.D., Moreau, L.: Provenance for Online Decision Making. In: Ludäscher, B., Plale, B. (eds.) Provenance and Annotation of Data and Processes, Lecture Notes in Computer Science, vol. 8628, pp. 44–55. Springer International Publishing (2015)