Provenance and Trust in Volunteered Geographic Information: The Case of OpenStreetMap

Carsten Keßler, Johannes Trame and Tomi Kauppinen

Institute for Geoinformatics, University of Muenster, Germany
carsten.kessler | johannestrame | tomi.kauppinen @uni-muenster.de

Abstract. We propose to use the trustworthiness of features in OpenStreetMap as a proxy function for data quality, where a feature's trust value is computed solely from its history. The trustworthiness is based on the different contribution patterns that can be found in a feature's history, such as rollbacks or deletions. We argue that these patterns influence the trustworthiness of a feature and its contributors' reputation. We discuss different potential implications of the patterns on trust and reputation.

Keywords. Data Quality, Provenance, Trust, Reputation, Volunteered Geographic Information, OpenStreetMap, Editing Patterns, Social Semantics

1 Introduction

Volunteered Geographic Information (VGI) [5] is increasingly attracting attention for professional use. The growing number of applications and projects building on OpenStreetMap (OSM) make data quality [4] an important issue. As traditional criteria for data quality do not apply for OSM, we propose to use the trustworthiness [9] of features in OpenStreetMap as a proxy function for data quality [3,8]. Similar to previous approaches developed for Wikipedia [1,7], the features' trust ratings are computed solely from their provenance. We look for specific patterns, such as rollbacks or deletions, that emerge when single features in OSM develop over time with input from different users. The trust ratings follow arguments about the patterns' implications for the trustworthiness of a feature and the contributors' reputation.

2 Provenance-based Trust and Reputation Model

OpenStreetMap stores a full copy of the current state of a feature when it is updated. In order to keep track of changes to the data, we have extended the provenance vocabulary introduced in [6] with classes and relationships specific to

OSM.[1] These annotations lay the foundations for structured analyses of emerging patterns in the creation of contents in OSM.

In order to get an overview of the different patterns that emerge during the collaborative generation of OSM features, we have developed a Web application [10] to visually explore their history, which led to the identification of the following four basic patterns. For each pattern, we discuss its expected effects on a feature's trustworthiness as well as the contributors' reputation. These effects need to be evaluated in future research:

- **Confirmations** refer to patterns where the existing data is (explicitly or implicitly) confirmed by other users. The underlying idea is that the more people have checked the information on a feature without changing it, the more likely it is that this information is correct (*many eyes principle*). Confirmations are hence edits where a user only adds data to a feature, or commits a changeset in the vicinity of the feature under consideration. The more confirmations apply to a feature, the more trustworthy it is and the higher the contributing users' reputation should be for adding reliable information.
- **Corrections** refer to edits that change a feature's geometry or tag description. We follow the assumption that more obvious errors are corrected more quickly than small errors in the geography of a feature, for example. The decrease function should hence be tied to the timespan between the faulty edit and the correction, i.e., the faster an error is corrected, the stronger the decrease in the first user's reputation. Corrections to a feature make it less persistent; however, they also indicate effort to improve. The effect on the feature's trustworthiness is hence subject to empirical evaluations.
- **Rollbacks** refer to corrections that revert a feature to a previous state. Such rollbacks[2] point to the fact that an update was faulty from the point of view of the user who made the rollback. A rollback is hence defined by three subsequent versions of a feature, where the first and last of the three subsequent versions of a feature are equal. Rollbacks should decrease the reputation of the user who submitted the second version, and eventually also (slightly) decrease the feature's trustworthiness for a loss in persistence.
- **Self-rollbacks** are special cases of rollbacks where users revert their own updates. These self-rollbacks[3] occur when a user notices her own mistakes after submission and corrects them. Further evaluations are required to make safe statements about the effect of self-rollbacks on trust and reputation, as the user made a mistake, but also made the effort to correct it.

We take a detailed perspective on the OSM features here by applying these measures to the *statements* that form a specific version of a feature. We decompose features into the smallest information units, i.e., into *triples* forming a Linked Data [2] graph. This approach uses the dualism that is created by combining the static view on a feature and the provenance view that highlights the

[1] See http://carsten.io/osm/osm-provenance.rdf.

[2] See, e.g., http://giv-heatmap.uni-muenster.de:4434/history/node/88875206

[3] See, e.g., http://giv-heatmap.uni-muenster.de:4434/history/node/368417050

changes between different versions. We can then assign a trust value $v \in [0, 1]$ to any of these statements in the static view, which allows users of OSM data to pose queries against the trustworthiness of the results, for example.

3 Conclusions and Outlook

Annotating OSM data with our provenance vocabulary allows us to make implicit information about the lineage of features in OpenStreetMap explicit. We have outlined how patterns in these annotated data can serve as input to a model that computes the trustworthiness of OSM features and the reputation of OSM contributors. The annotation and pattern extraction has already been implemented. We are currently evaluating the approach using a OSM history dump for the city of Berlin, Germany.

Acknowledgments

Funding came from the Int. Research Training Group *Semantic Integration of Geospatial Information* (DFG GRK 1498; http://irtg-sigi.uni-muenster.de) and the *SimCat II* project (DFG JA1709/2-2; http://sim-dl.sf.net).

References

1. B. Adler and L. De Alfaro. A content-driven reputation system for the Wikipedia. In *Proc. 16th int. conference on World Wide Web*, pages 261–270. ACM, 2007.
2. T. Berners-Lee. Linked Data. Personal view available from http://www.w3.org/DesignIssues/LinkedData.html, 2009.
3. M. Bishr and W. Kuhn. Geospatial Information Bottom-Up: A Matter of Trust and Semantics. In S. I. Fabrikant and M. Wachowicz, editors, *The European Information Society – Leading the Way with Geo-information*, Lecture Notes in Geoinformation and Cartography, pages 365–387. Springer-Verlag Berlin Heidelberg, 2007.
4. N. Chrisman. The error component in spatial data. *Geographical information systems*, 1:165–174, 1991.
5. M. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, August 2007.
6. O. Hartig. Provenance Information in the Web of Data. In *Proceedings of the Linked Data on the Web (LDOW) Workshop at the World Wide Web Conference (WWW), Madrid, Spain*, April 2009.
7. S. Javanmardi and C. Lopes. Modeling trust in collaborative information systems. In *International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2007)*, pages 299–302. IEEE, 2008.
8. C. Keßler, K. Janowicz, and M. Bishr. An Agenda for the Next Generation Gazetteer: Geographic Information Contribution and Retrieval. In *GIS '09: Proc. 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Seattle, Washington*, pages 91–100, New York, 2009.
9. P. Sztompka. *Trust: A sociological theory*. Cabridge University Press, 1999.
10. J. Trame and C. Keßler. Exploring the Lineage of Volunteered Geographic Information with Heat Maps. In *GeoViz 2011, Hamburg, Germany*, 2011.