Assessing OSM street semantics quality using context and user trustworthiness

Amerah Alghanim and Michela Bertolotto

School of Computer Science
University College Dublin, Ireland
Emails: amerah.alghanim@ucdconnect.ie
michela.bertolotto@ucd.ie

Abstract

OpenStreetMap (OSM) provides free tools that anybody can use to edit maps and upload spatial data. This led a strong interest in investigating the usability of this data in terms of its limitation and potential. The main concerns relate to OSM data quality. This paper focuses on analysing and evaluating the quality of the OSM street network tags based on specific features of streets, their context and the trustworthiness of users who have edited those tags. In our work, spatial data mining techniques are applied for extraction of OSM features. Machine learning techniques are used for automatically evaluating the data extracted and predict the street type. Advanced validation techniques are used to assess the OSM contributors' trustworthiness.

Keywords: OpenStreetMap, Spatial Data Mining, Machine Learning, User Trustworthiness, Data Quality, Street Network Analysis.

1. Introduction

OSM is often referred to as the Wikipedia of maps of the world (Arsanjani et al., 2015; Haklay and Weber, 2008). OSM provides free tools that anybody can use to edit maps and upload spatial data. This ability directed to freely available geographic information to grow, but since the data is edited by volunteers there are concerns related to OSM data quality. Spatial data quality can be studied based on several dimensions identified including attribute accuracy, geometry accuracy, completeness, logical consistency, lineage, temporal accuracy, semantic accuracy and user trustworthiness (see (Jilani et al., 2017) for an overview).

The main focus in this paper is analysing and evaluating the quality of the OSM street network tags based on specific features of streets as well as their context and user trustworthiness. We apply spatial data mining techniques for extraction of OSM features for different types of streets as well as machine learning techniques for automatically evaluating the data extracted and predict the street type. Specific features we consider that have not yet been extensively studied are the context of a street (for example, the types of buildings around specific types of streets, etc.) and the trustworthiness of users that inputted and edited the street.

As volunteers are the primary contributors in the input of OSM data there are concerns regarding their trustworthiness and credibility of the data they provided. To evaluate the quality of OSM data, advanced validation techniques are needed to assess the OSM contributors' trustworthiness. The approach we are developing uses trust and reputation of contributors based on spatial and temporal dimensions.

In the remainder of the paper we describe our novel approach to automatically assess the quality of OSM street tags based on a combination of specific street features combined with a measure of the trustworthiness of users that edited those streets. Our approach can also be used to predict the street type for streets that do not have associated tags.

2. Related Work

As OSM data is contributed by users with varying levels of knowledge and practice, not all contributed data has the same quality. Therefore, a lot of research has been carried out to assess such a quality. For example, in (Kashian et al., 2016), a tool to analyse and discover meaningful patterns among recorded objects in OSM with a focus on POI (points of Interest) has been introduced. In (Ali et al., 2014) a classification-based approach to manage and improve VGI data was proposed, while in (Brovelli et al., 2016) an automated methodology to compare OSM with authoritative road datasets in terms of spatial accuracy was developed. In (Jilani et al., 2017) machine learning was used for automatically predicting the semantics (tags) of the street. While promising results were achieved, the approach presented some limitations as, for example, it did not consider the relationship between streets and the objects surrounding them.

The effect of user trustworthiness on the quality of the data they contribute has also been studied. For example, in (Mooney and Corcoran, 2012) a method to access the OSM history data and studied some contributors characteristic by applying social network analysis techniques was designed. In (Bishr and Werner, 2013) the authors developed a computational model to use trust and reputation of contributors with spatial and temporal dimensions for quality assessment. The main aspect they identified in addition to previous research is that of Informational Trust. They define Informational Trust as an object that can be inferred by interpersonal trust between the originator and the consumer. For example, in the case of VGI the trust on the data contributed by the volunteers is used as informational trust. In our research we follow this approach and adapt it to the assessment of contributor trustworthiness in the context of street networks semantics in OSM.

3. Approach

Our approach to extracting street features relies on spatial data mining and statistical techniques. In particular, statistical search queries were applied to find the value of specific attributes of streets such as length, max speed, etc. that are useful to identify their type. We also extracted the context of street and analysed, for example, the type of buildings that typically surround a certain type of street. Our aim is to apply machine learning techniques to predict the value of each street type according to the context and features of streets extracted.

An important feature in our work relates to the trustworthiness of contributors. In order to evaluate it we extracted the history of OSM edits. Following the user trustworthiness model proposed in (Bishr and Werner, 2013), our approach sums trust and reputation of volunteers on spatial and temporal dimensions. Our case study relates to the street network of the city of London (UK). In our approach, the trustworthiness of users is impacted by the edits that other volunteers make on objects contributed by them and takes into account their spatial proximity the position of the objects edited (e.g., distance from provenance of volunteer to location of objects edited). As stated in (Bishr and Werner, 2013), it is generally assumed that a person trying to maintain a good reputation will try to provide high quality information. Therefore, the concept of *Information Trust* is used as a proxy for user trustworthiness and is defined as:

$$\overline{T}_{ij} = T_{ij} \,. K_j$$

Equation 1 Information Trust

Where T_{ij} represents the information trust of volunteer *i* at moment *j* in his/her history. A time decay factor K_j is used in this model as we assume that information trust is subject to decay over time (the information provided will potentially not be accurate/up-to-date after a long time). K_j is a function of the time passed from the moment *j* when the objects were updated.

In adapting this model to our case study, we consider an update o_{ik} made by volunteer *i* on an object *o* at time *k.* An associated quality value $q(o_{ik})$ is calculated based on edits made by others on the same object (and whether those edits were reverted). The quality of edits made by volunteers affect their reputation and information trust. We will consider specific threshold values for calculating quality. The value and weight of the different parameters involved will be determined empirically.

4. Conclusion

In this paper we discussed our assessment approach for OSM data that relies on spatial data mining and statistical techniques. The approach extracts specific street features that are useful to identify the street type. Among these we consider a street context (e.g., the types of buildings that typically surround a certain type of street). We also adapted an existing model to measure user trustworthiness which takes into account both spatial and temporal aspects. In the future, we will evaluate our approach by assessing not only the street networks of London but also those of other cities (e.g., Dublin) as well as those of less populate areas where OSM coverage is less detailed. The application of machine learning techniques, including Naïve Bayes, nearest neighbour and decision trees, for assessing the OSM street network semantics produced promising results (Jilani et al., 2017). However, such work did not take street context and user trustworthiness into account. We think the inclusion of these features will provide improved results. The ultimate goal of our research is the development of a web-based tool that will help contributors associated tags to street in OSM.

5. Acknowledgements

This work was supported by the Saudi Arabian Bureau (Ministry of Education) through the King of Saudi Arabia Scholarship Program.

6. References

Ali, A., Alsalman, R., Schmid, F. & Kauppinen, T., 2014. *Ambiguity and Plausibility: Managing Classification Quality in Volunteered Geographic Information.* Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, USA.

Arsanjani, J. J., Zipf, A., Mooney, P. & Helbich, M., 2015. *An Introduction to OpenStreetMap in Geographic Information Science: Experiences, Research, and Applications.*Springer International Publishing.

Bishr, M. & Werner, K., 2013. *Trust and Reputation Models for Quality Assessment of Human Sensor Observations.*Spatial Information Theory. COSIT 2013. Lecture Notes in Computer Science, vol 8116, Springer.

Brovelli, M., Minghini, M., Molinari, M. & Mooney, P., 2016. *Towards an automated comparison of OpenStreetMap with authoritative road datasets.* Transaction in GIS. John Wiley & son.

Haklay, M. & Weber, P., 2008. *OpenStreetMap: User-Generated Street Maps.* IEEE Pervasive Computing, vol. 7, Issue 4.

Jilani, M., Bertolotto, M., Corcoran, P. & Alghanim, A., 2017. Traditional vs. Machine Learning Techniques for OSM quality assessment. In: *Volunteered Geographic Information and the Future of Geospatial Data.*IGI Global.

Kashian, A., Richter, K.-F., Rajabifard, A. & Chen, Y., 2016. *Mining the Co-existence of POIs in OpenStreetMap for Faulty Entry Detection.* Proceedings of the 3rd Annual Conference of Research@Locate.

Mooney, P. & Corcoran, P., 2012. *Who are the contributors to OpenStreetMap and what do they do?.* Lancaster University, UK.