The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-3/W3, 2015
ISPRS Geospatial Week 2015, 28 Sep – 03 Oct 2015, La Grande Motte, France

TRUSTING CROWDSOURCED GEOSPATIAL SEMANTICS

P. Goodhue, H. McNair , F. Reitsma

Dept. of Geography, University of Canterbury, Christchurch, New Zealand
(paul.goodhue, hamish.mcnair)@pg.canterbury.ac.nz, femke.reitsma@canterbury.ac.nz

KEY WORDS: Crowdsourcing, Semantics, Ontology, Data quality, Trust

ABSTRACT:

The degree of trust one can place in information is one of the foremost limitations of crowdsourced geospatial information. As with the development of web technologies, the increased prevalence of semantics associated with geospatial information has increased accessibility and functionality. Semantics also provides an opportunity to extend indicators of trust for crowdsourced geospatial information that have largely focused on spatio-temporal and social aspects of that information. Comparing a feature's intrinsic and extrinsic properties to associated ontologies provides a means of semantically assessing the trustworthiness of crowdsourced geospatial information. The application of this approach to unconstrained semantic submissions then allows for a detailed assessment of the trust of these features whilst maintaining the descriptive thoroughness this mode of information submission affords. The resulting trust rating then becomes an attribute of the feature, providing not only an indication as to the trustworthiness of a specific feature but is able to be aggregated across multiple features to illustrate the overall trustworthiness of a dataset.

1. INTRODUCTION

Crowdsourced geospatial information will transform data collection and its legitimacy is contingent on the trustworthiness of the information submitted. We can assess this from a range of perspectives: the spatio-temporal nature of the information, the social aspect of the information (e.g. the provider of the information), and the semantics of the information. In this paper we focus on the semantics of crowdsourced geospatial information, and how we can define a measure of trust using these semantics. Assessing the trust of the semantics associated with crowdsourced geospatial information ensures we know to what degree it is useful. This appraisal can be performed via comparisons of the intrinsic and extrinsic properties of geospatial features to ontologies to produce a feature level trust rating. This trust rating can then be used as an attribute to aid in discovery and analysis of the information, and be aggregated across multiple features to provide an indication of trustworthiness for the data set as whole.

2. RELATED WORK

Research on the semantics of geospatial information find it supports better discovery (Egenhofer 2002), use (Reeve and Han 2005), and access (Yue et al. 2007, Janowicz et al. 2012). Today the increase in use of semantics for search and discovery has led to a proliferation of websites that use ontologies in their backend, e.g. ImageNet (Deng et al. 2009). Tie this to collectively developed tags or annotations (Marcheggiani et al. 2007), and the potential for crowdsourced geospatial datasets being integrated into all stages of the spatial data supply chain is substantial.

Using the crowd involves significant cognitive diversity. To ensure that the semantics expressed in the attributes associated with crowdsourced geospatial features are accurate, we can either leverage this cognitive diversity and apply statistical methods for measuring the agreement of concepts thereby embracing the diversity of responses (Narock and Hitzler 2013), or constrain user input. Semantics can be used to determine the difference among concepts in geospatial datasets (Kuhn 2005). These differences may be subtle variations in class, e.g. a Munro is a Mountain in some instances, to more distinct differences such as a Coconut Palm and a Date Palm. Measures of semantic similarity have been developed to ascertain how close concepts are to each other (e.g. Janowicz et al. 2011, Sizov 2010), and can be used to compare ontologies and improve spatial datasets (Ballatore et al. 2013).

Measures of trust for crowdsourced geospatial information usually consider the spatial and social aspects of the crowdsourced geospatial information (Goodchild and Li, 2012), but often overlook the semantic aspect. Measures of spatio-temporal trust of crowdsourced geospatial information determine the spatial accuracy of the information (Haklay 2010), such as its shape, orientation and location. Social measures of the trust of crowdsourced geospatial information are based on the reputations of the information producers (Bishr and Janowicz, 2010). For example, the USGS National Mapping Corps project uses the reputation of information producers to determine trust of the information (McCartney et al. 2013).

Some consideration has been given to certain aspects of the semantics of crowdsourced geospatial information, such as the work by Vandecasteele and Devillers (2013), who proposed the use of semantic similarity to constrain the crowdsourced information by notifying the producer of the information when two attributes were too similar or dissimilar. The concept of semantic similarity can be used at the feature level (i.e. on individual pieces of information) to measure the trust of crowdsourced geospatial information by comparing the feature to trusted features of a similar type, both within the crowdsourced dataset or with external datasets (Ramos et al. 2013). Likewise Bordogna et al. (2014) semantically assessed the quality of crowdsourced geospatial information through a linguistic approach using a hierarchical structure of attribute categories (e.g. tags) within the information. Features given a general category are deemed to be less accurate than features given a specific category, and feature assigned multiple

This contribution has been peer-reviewed.
Editors: A.-M. Olteanu-Raimond, C. de-Runz, and R. Devillers
doi:10.5194/isprsarchives-XL-3-W3-25-2015

categories are also deemed inaccurate as this shows uncertainty in the producer's ability to categorise the feature. We build on these methods with further methods for using semantics to assess the trustworthiness of crowdsourced geospatial information.

3. CROWDSOURCED SEMANTICS

How do we trust the semantics of information provided by the crowd? Improving the knowledge of the semantic quality of the information is achieved through assessing or constraining the semantic aspect of the information. Constraining the semantic aspect of the information can improve the quality of the information but can limit the knowledge captured with the information, therefore semantically unconstrained information coupled with semantic quality assessments can provide us with diverse and trustworthy geospatial information. Semantic quality assessments of crowdsourced geospatial information can be performed on the intrinsic or extrinsic semantics of the information, where the intrinsic semantics involves the feature and its consistency both internally and with linked ontologies, and where extrinsic semantics investigate the feature and its consistency within a wider context of related information and ontologies.

A theoretical example of a crowdsourcing project that would benefit from improved semantic trust is a project that crowdsources the location of fruit producing trees on public land, and some information that describes the trees. The crowdsourced geospatial information would include a point feature depicting the trees location and information about the tree that would describe the type of tree, the quality and quantity of the fruit on the tree, accessibility of the tree (e.g. across a stream or up a bank) and other observable attributes of the tree such as height and diameter. This information would be linked to an ontology describing fruit producing trees and other external data and ontologies to help to determine the semantic trust of the information. Although projects exist that crowdsource fruit tree information (e.g. www.ediblecities.org), little is done with regards to determining the trust of the information, especially the semantic trust of the information. By determining the semantic trust of the crowdsourced information, anyone wishing to use the information can trust that the information accurately describes a tree in the real world.

Intrinsic semantic assessments of crowdsourced geospatial information assess the feature within the context of its containing geospatial dataset. Intrinsic assessments are based on comparing the feature type and attributes with an ontology. Through intrinsic semantic assessments we can determine the internal consistency of the feature and whether or not it conforms to the semantics of its related ontology. A feature that is internally consistent and conforms to an ontology based on its attributes and feature type would be assigned a high rating of trust for its intrinsic semantic component. For example, a crowdsourced fruit tree feature could contain an attribute describing the type of tree the feature represents. Through assessments of this attribute and a tree ontology we could determine if the feature the producer has created describes a fruit producing tree in the real world, as outlined in figure 1. If a producer creates a tree feature and describes its type as "Apple", the tree ontology would tell us that "Apple" is a type of fruit producing tree and therefor the features type is consistent with fruit trees and is somewhat trustworthy. Alternatively, if the producer was to submit a tree feature with a type of "Rose", we would assign this feature a low trust rating because although "Rose" is a type of plant, it would not fit into a subclass of fruit trees.

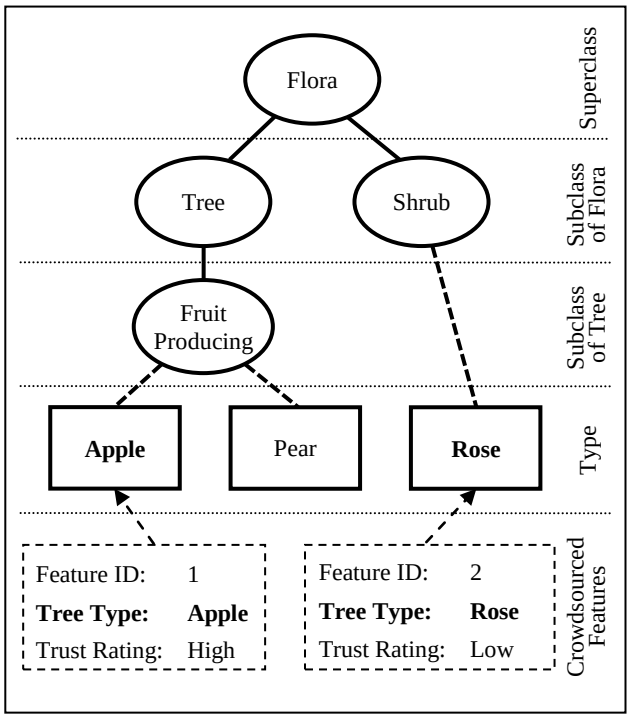

Figure 1. Assessing crowdsourced features with an ontology to determine if their Tree Type attribute is a type of fruit tree.

Each feature is associated with attributes. These attributes may be assessed based on the semantics of their data type or object type. For example, a point feature representing a fruit tree in our example crowdsourcing project may include attribute data with a tree type of "Coconut" and with a height of 100m. Comparing our geospatial dataset against a tree ontology, we would find that Coconut trees are a type of fruit producing tree but do not usually grow taller than 30m. We would assign a low trust rating to this information because the attribute data is inconsistent and therefore untrustworthy. Object properties may also be evaluated for trustworthiness. A feature representing a single fruit tree could not have a geometry type of "Line", and if a member of the crowd created such a feature, this information would also be deemed untrustworthy.

Semantic trust can also be measured through extrinsic assessments, that is, comparing the crowd created information with external datasets. Extrinsic semantic assessments measure the trust of crowdsourced geospatial information by determining how a feature fits into its spatio-temporal surroundings by comparing the semantics of the feature to the semantics of its surroundings. For example, if a point feature was created as part of our example project of a Coconut tree in Antarctica today, intrinsically the information could be trustworthy as the Coconut tree feature may have the appropriate attributes, however comparing our piece of crowdsourced information against other ontologies regarding the climatic conditions Coconut trees grow in, the location of that point feature and the climate at that location, we would not trust this information. Taking into account the spatial and temporal characteristics is important for incorporating contextual knowledge about the crowdsourced information. Adding this additional level of trust

to the crowdsourced information makes the information more usable and helps to define the line between the information being completely untrustworthy (e.g. the fruit tree does not exist in the real world) or somewhat untrustworthy but still usable (e.g. the fruit tree exists but is not fruiting at the time of year the information states that it is). Extrinsic semantic assessments can become complex as they do not focus solely on the feature and its meaning, but require descriptions of the ontological relationships between the feature and the wider environment.

Methods for assessing the semantic quality of crowdsourced geospatial information become more complex as the information becomes thematically richer. In cases where the crowdsourced information is constrained, the attribution is common throughout the dataset and the features are likely to be both internally consistent and semantically similar to other features. But by semantically constraining crowdsourced information we risk losing knowledge of the feature that the producer could have otherwise supplied with the feature. Measures of semantic trust are needed for semantically unconstrained crowdsourced information to leverage the diversity and cognition of the crowd while maintaining trust in the crowdsourced information. Alongside measures of the semantic trust of crowdsourced geospatial information are measures of the trust of its spatio-temporal and social components. Intrinsic and extrinsic semantic assessments of crowdsourced geospatial information forms the semantic component of a larger crowdsourcing model that also encapsulates assessments of the intrinsic and extrinsic aspects of the spatio-temporal and social components of the crowdsourced geospatial information. Although measuring the semantic trust of crowdsourced geospatial information provides us with a rating of trust of the information, measures of the spatio-temporal and social trust of the information help to provide an overall trust rating for the information. The wider crowdsourcing model that the semantic assessments form a part of assesses both the crowdsourced information itself and how it fits into its surroundings. The author is also considered in order to determine their influence on the trust of the information. By focussing on the information itself, the semantic and spatio-temporal components of the crowdsourcing model complement each other and the use of both helps to improve the trust of the information further by generating trust ratings that represent all aspects of the information itself.

4. DIRECTION OF RESEARCH

Previous applications of trust models for spatial datasets (Malaverri et al., 2012) and crowdsourced geographic features (Bishr & Mantelas , 2008; Celino, 2013) have produced a scalar value as a proxy for trustworthiness. This can then be employed in much the same way as metadata or provenance information to give an indication of data quality. However, aside from just determining trustworthy and untrustworthy data the use of a single metric provides several notable advantages over traditional, text based metadata. Firstly, for those unfamiliar with the use of traditional datasets (such as members of the crowd) it provides a simple indication as to whether a dataset or feature is appropriate without having to understand the intricacies of metadata documentation. Further to this, it provides scope for the aggregation of feature level trust ratings to give a representation of trustworthiness for a dataset as a whole, as outlined in figure 2. As crowdsourced datasets often have multiple and varied sources this numeric approach is simpler to implement and understand than attempting to aggregate the individual metadata records of each feature into a coherent article. Finally, quantifying trust not only makes for a convenient feature attribute, but this attribute – being a number as opposed to text – is easily incorporated into computerised systems, giving the data a degree of self-description.

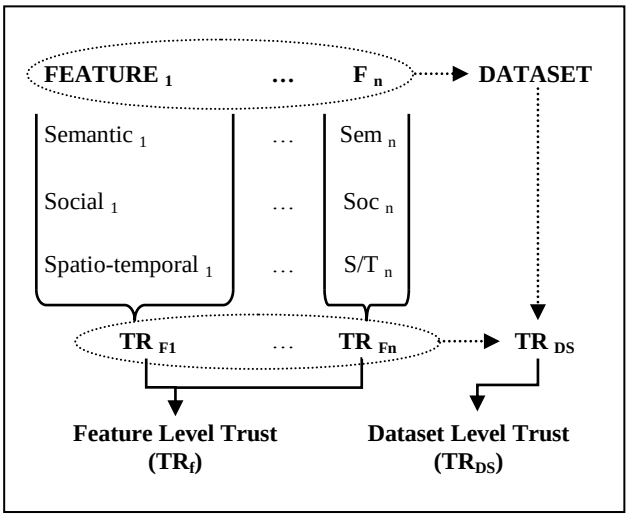

Figure 2. Formation of trust ratings corresponding to feature (TR_F) and dataset (TR_{DS}) levels; in the same manner as a dataset is an accumulation of features a datasets trust rating is an aggregate of feature trust ratings.

Self-describing data increases functionality by providing improved search and implementation capabilities. For example, if a user requires a map of the most trustworthy of fruit tree features (in, say, the interest of maximising time spent collecting fruit versus finding trees) they can perform a simple attribute based search – features with high trust ratings. Alternatively, to enable the utilisation of an entire dataset, a feature's influence can be made proportional to its trust rating. If one were to analyse distribution of fruit trees (to, say, evaluate the impact of birds dispersing seed as opposed to gardeners purchasing plants) the model could be structured to increase the influence of features in the analysis based on their trust ratings. This approach is similar to spatial analysis techniques where the weighting of inputs is based on the perceived quality of the data from which they were derived (MacCormack & Eyles, 2010).

An additional strength of numeric trust ratings is in their ability to aid in the calibration and strengthening of the models which produce them. Comparison of crowdsourced datasets with their authoritative counterparts have been used to prove accuracy and completeness (Haklay, 2010). However, by comparing the analysis of a crowdsourced dataset (via the aforementioned weighted approach) with equivalent analysis of an authoritative dataset it is possible to identify under which scenarios various aspects of the trust model should project the greatest level of influence. Such an approach allows for not only the determining of the importance of the intrinsic and extrinsic assessments of semantic trust, but also the role of semantic trust within the wider trust model as a function of the dataset and what it represents. Such calibration provides for a more robust assessment and subsequently improving overall accuracy and functionality of trust models and the ratings they produce.

5. CONCLUSION

Crowdsourced geospatial information makes use of the diverse knowledge of the crowd, but for the information to be useful it must be trusted. Measures of the trust of crowdsourced geospatial information focus on different aspects of the information, such as the spatio-temporal, social and semantic aspects. The trust of the semantic aspect of crowdsourced geospatial information is often overlooked by crowdsourcing applications, but in unconstrained data models it can strengthen the understanding of the overall trust of the information. Through intrinsic and extrinsic semantic assessments of the quality of crowdsourced geospatial information we can measure the semantic trust of the information, and by coupling these assessments with a wider trust model of the spatio-temporal and social aspects, we can improve the overall trust of the information.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the CRCSI, project 3.02.

REFERENCES

Ballatore, A., Bertolotto, M., & Wilson, D. C. (2013). Geographic knowledge extraction and semantic similarity in OpenStreetMap. Knowledge and Information Systems, 37(1), pp.61-81.

Bishr, M., & Janowicz, K. (2010). Can we trust information? - the case of volunteered geographic information. CEUR Workshop Proceedings, 640.

Bishr, M., & Mantelas, L. (2008). A trust and reputation model for filtering and classifying knowledge about urban growth. GeoJournal, 72(3-4), 229-237.

Bordogna, G., Carrara, P., Criscuolo, L., Pepe, M., & Rampini, A. (2014). A linguistic decision making approach to assess the quality of volunteer geographic information for citizen science. Information Sciences, 258, pp.312-327.

Celino, I. (2013). Human computation VGI provenance: semantic web-based representation and publishing. Geoscience and Remote Sensing, IEEE Transactions on, 51(11), 5137-5144.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database, Proc. CVPR 2009, IEEE (2009), pp.248-255.

Egenhofer, M J (2002). Toward the semantic geospatial web. In Proceedings of the 10th ACM international symposium on Advances in geographic information systems (GIS '02). ACM, New York, NY, USA, 1-4.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. Spatial Statistics, 1, pp.110-120.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and Planning B: Planning and Design, 37, pp.682-703.

Janowicz K, M. Raubal, and W. Kuhn (2011). The semantics of similarity in geographic information retrieval, Journal of Spatial Information Science, 2, pp.29-57.

Janowicz K, S. Scheider, T. Pehle, and G. Hart (2012). Geospatial Semantics and Linked Spatiotemporal Data – Past, Present, and Future. Semantic Web 0 1–0. 1. IOS Press.

Kuhn, W. (2005). Geospatial Semantics: Why, of What, and How? Journal on Data Semantics 3, pp.1-24.

MacCormack, K. E., & Eyles, C. H. (2010). Enhancing the Reliability of 3D Subsurface Models through Differential Weighting and Mathematical Recombination of Variable Quality Data. Transactions in GIS, 14(4), 401-420.

Malaverri, J. E., Medeiros, C. B., & Lamparelli, R. C. (2012, March). A provenance approach to assess the quality of geospatial data. In Proceedings of the 27th Annual ACM Symposium on Applied Computing (pp. 2043-2044). ACM.

Marcheggiani, E., Nucci, M., Tummarello, G., & Morbidoni, C. (2007). Geo Semantic Web Communities for Rational Use of Landscape Resources. Ontologies for Urban development: Conceptual Models for Practitioners, Turin, Italy, pp.100-113.

McCartney, E., Bearden, M., & Newell, M. (2013). Crowd-Sourcing the Nation: Now a National Effort. Retrieved from http://www.usgs.gov/newsroom/article.asp?ID=3664#.VP5pNfmUc40

Narock, T., & Hitzler, P. (2013). Crowdsourcing Semantics for Big Data in Geoscience Applications. Semantics for Big Data: Papers from the AAAI Symposium. Presented at the AAAI 2013 Fall Symposium Series Semantics for Big Data, Arlington, VA, November 15-17, 2013.

Ramos, J. M., Vandecasteele, A., & Devillers, R. (2013). Semantic Integration of Authoritative and Volunteered Geographic Information (VGI) using Ontologies. Association of Geographic Information Laboratories for Europe (AGILE).

Reeve L, and Han, H. (2005). Survey of semantic annotation platforms. In Proceedings of the 2005 ACM symposium on Applied computing, ACM, pp.1634-1638.

Sizov S (2010). GeoFolk: Latent Spatial Semantics in Web 2.0 Social Media, in: Proceedings Web Search and Data Mining, 2010.

Vandecasteele, A., & Devillers, R. (2013). Improving volunteered geographic data quality using semantic similarity measurements. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 1(1), pp.143-148.

Yue, P., Di, L., Yang, W., Yu, G., & Zhao, P. (2007). Semantics-based automatic composition of geospatial Web service chains. Computers & Geosciences, 33(5), pp.649-665.