Automated Assessment and Improvement of OpenStreetMap Data

Musfira Jilani[1], Padraig Corcoran[1], Michela Bertolotto[1]

[1]School of Computer Science and Informatics, University College Dublin, Ireland
Tel. (+353 1 7162483) Fax (+353 1 7162469)
Email: musfira.jilani@ucdconnect.ie

KEYWORDS: OpenStreetMap, Topological Spatial Relations, Multivariate Bernoulli Distribution

ABSTRACT: The challenging task of creating an accurate and up-to-date map of the fast changing world can be easily solved by means of crowdsourcing. OpenStreetMap (OSM) is a good example of this approach. However, the quantification of quality and suitability of OSM for certain applications is necessary as this database is primarily constructed by amateurs. Researchers have proposed several parameters to assess the quality of OSM. It is found that most of the current research in this area focusses on geometric accuracy and is confined to road networks. This paper analyses the logical and semantic accuracy of OSM and includes other geographic features along with street network.

1. Introduction

With the advent of Web 2.0 there has been a rapid increase in the generation of User Generated Content (UGC). From serious research to gossip, the accessibility and affordability of communication technology has made UGC ubiquitous. OpenStreetMap (OSM) popularly referred to as “Wikification of Geographic Information Sciences (GIS)” (Sui 2008) is a good example of UGC where citizens contribute in the production of geographic information.

Founded by Steve Coast in July 2004 in London, OSM is a collaborative project that aims to provide a freely accessible and editable map of the world in raster and vector formats. This is accomplished by means of crowdsourcing i.e., the task of generating the geographic database of the world is outsourced to the common people located across the world. Citizens equipped with GPS mobile devices act as sensors (Goodchild 2007) and help in the mapping of very fine details which would otherwise have been impossible. This information is modelled as a triple (resource, property, value) as defined in the Resource Development Framework (RDF) model (Manola and Miller 2004). Resource represents the coordinates of a geometric primitive say, a point. Property corresponds to the tag associated with the resource; although commonly used tags such as ‘highway’, ‘motorway’ etc. have been provided for convenience, albeit a new tag suggested by the users can also be accommodated after a vote. Value contains information about the given tag. The RDF information so generated is made available in XML Syntax for free download and use under Creative Commons Attribution Share Alike 2.0 license.

Since its inception the number of registered contributors to OSM is increasing at a fast pace and so is the volume of the generated geographic information. This database is increasingly finding several useful applications. Popular services such as Apple Inc., Wikipedia, Flickr, Foursquare etc., are using OSM to provide some sort of map rendering. Several mapping projects with specific interests are being undertaken. Some of these include projects targeting the mapping of border and administrative boundaries, marine, shipping, environment, humanitarian relief, and accessibility routes for the visually challenged. This plethora of crucial applications necessitates the validation of accuracy of OSM data.

Several elements to assess the quality of geographic databases were proposed by Girres and Touya (2010). These include geometric, semantic, logical and temporal accuracy, logical consistency, completeness, usage and lineage. Mashhadi et al. (2013) found that contextual factors such as the prevailing socio-economic conditions coupled with population density determine the completeness of OSM information in a particular region. Helbich et al. (2012) observed that spatial heterogeneity exists in positional accuracy with clusters of high and low accuracies.

However, a limitation to most of the studies on OSM is their focus on road networks. According to Mulligann et. al (2011), "Most of these (OSM) studies do not take geographic feature types into account. This is for two reasons. First, and in contrast to street networks, choosing a reference set is difficult. Second, we lack the measures to quantify the degree of feature type mis-categorization." This paper presents a novel approach to assess the inconsistencies in OSM database with respect to several geographic features including the road networks. It does this by quantifying the logical consistency and semantic accuracy of the OSM database in terms of topological spatial relationships existing between several objects such as buildings, roads, water bodies, Points-of-Interests (POIs). A statistical machine learning approach is used to learn these relationships and observe inconsistencies in the dataset.

The rest of the paper is divided as follows. Section 2 describes the dataset used in the study. Section 3 discusses the methodology adopted to determine the semantic accuracy within the chosen dataset. This is followed by Section 4 where the inferences made from this study are discussed.

2. Dataset

OpenStreetMap database for Ireland, Europe's third largest island has been chosen for this study. The initial analysis is limited to the topological relationship between houses and roads. For this purpose Maynooth, a small university town towards the north of Ireland, having a good number of residential buildings in proximity to road network, is chosen. A dataset comprising of 100 houses surrounded by three roads spread over a 1 kilometer square region in Maynooth has been extracted on the basis of associated semantic information and is studied.

3. Methodology

The approach followed to determine the semantic accuracy of the given dataset on the basis of topological relationships between geographic features is illustrated by means of a flowchart in Figure 1.

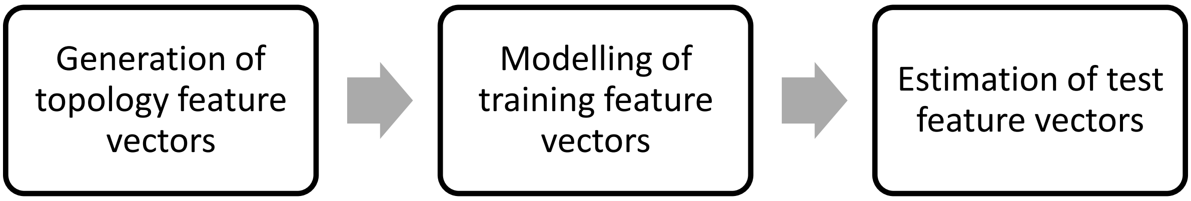
Generation of topology feature vectors
Modelling of training feature vectors
Estimation of test feature vectors

Figure 1: Flowchart describing the steps followed in determining semantic accuracy

In order to generate the topology feature vector, the Dimensionally Extended 9 Intersection Model (DE-9IM) proposed by Clementini et al. (1993) based on the works of Egenhofer and Robert (1991) is used. This model defines pair-wise spatial relationships between geometries of different types as pair-wise intersections of their interior, boundary and exterior and presents them in a 3 x 3 matrix. This is converted into a 9 x 1 feature vector. It is observed that a more useful feature vector *k* with lower dimensions such as $k=[k_1 , k_2 , \dots, k_N]^t$ where $N<9$ can be generated by overlooking some intersections like the intersection between the exteriors of houses and roads in our example. Figure 2 shows a useful feature vector generated corresponding to the topological relationship between a house and nearby road. This topology feature vector is binary in nature.

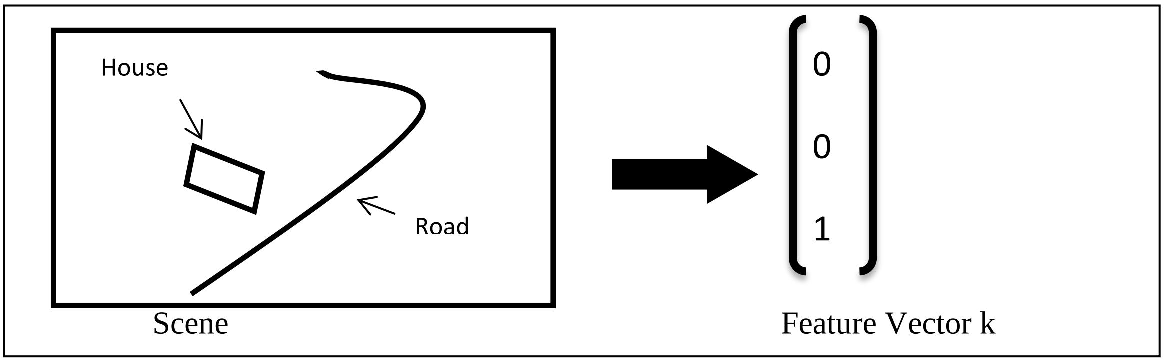

Figure 2: Feature vector representing the topological relationship shown in the scene.

Following the approach of unsupervised learning (Barber 2011) and assuming that most of the topological relationships in the existing OSM database are correct, 70 percent of the dataset is chosen for training and the remaining is used for testing. Multivariate Bernoulli distribution is used to learn the topology between a house and a road using the set of training feature vectors and to estimate the distribution of test feature vector. A slightly simplified version of the approach followed by Larochelle et al. (2010) and expressed in Equation 1 is followed.

$$p(k|p_1 p_2 \dots p_N) = \prod_{n=1}^{N} {p_n}^{k_n} \, (1 - p_n)^{(1-k_n)} \tag{1}$$

The estimation of the N parameters $(p_1 p_2 \dots p_N)$ in the above equation is done using the Maximum Likelihood Estimation. If the estimated parameters lie above a threshold, the corresponding house and road is said to be correct and the semantics associated with that particular house are considered accurate.

4. Discussion

An evaluation of this approach is currently under progress the results of which will be presented at the conference. It is envisaged that a good assessment of OSM quality can be made by determining the semantic accuracy of the data on the basis of associated topological information. The removal or correction of the detected inconsistencies can improve the database. Apart from a statistical machine learning approach to learn the topological relationships and estimate the inconsistencies, a rule-based learning approach can also be used to solve this problem.

Acknowledgements

This work is funded by Irish Research Council under the Embark Postgraduate Scholarship Scheme. The author gratefully acknowledges this support.

References

Clementini E, Felice P D and Oosterom P 1993 A Small Set of Formal Topological Relationships Suitable for End-user Interaction. *Advances in Spatial Databases* Springer Berlin/Heidelberg 277-295

Barber D 2011 *Bayesian Reasoning and Machine Learning*. Cambridge University Press

Egenhofer M J and Robert D F 1991 Point-set Topological Spatial Relations. *International Journal of Geographical Information System* 5(2), 161-174

Girres J-F and Touya G 2010 Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS* 14(4): 435-459

Goodchild M 2007 Citizens as Sensors: the World of Volunteered Geography. *GeoJournal* 69(4): 211-221

Helbich M, Amelunxen C and Neis P 2012 Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Propreitary Geodata. *International GI_Forum 2012* Salzburg Austria

Larochelle H, Bengio Y and Turian J 2010 Tractable Multivariate Binary Density Estimation and the Restricted Boltzmann Forest. *Neural Computation* 22(9):2285-2307

Manola F and Miller E 2004 RDF Primer, W3C Recommendation. WWW document, http://www.w3.org/TR/rdf-primer

Mashhadi A, Quattrone G and Capra L 2013 Putting Ubiquitous Crowd-sourcing into Context. *Proc. Of the 16th ACM International Conference on the Computer Supported Work and Social Computing (CSCW2013)*

Mulligann C, Janowicz K, Ye M and Lee W-C 2011 Analyzing the Spatial-semantics Interaction of Points of Interest in Volunteered Geographic Information. *Spatial Information Theory* 350-370

Sui D Z 2008 The Wikification of GIS and its Consequences: Or Angelina Jolie’s New Tattoo and the Future of GIS. *Computers, Environment and Urban Systems* 32: 1-5

Biography

Musfira Jilani is a first year PhD student in the School of Computer Science and Informatics at UCD. She has recently moved to Ireland after doing her undergraduate studies from India.

Padraig Corcoran is a post-doctoral researcher in the School of Computer Science and Informatics at UCD. Michela Bertolotto is a senior lecturer in the School of Computer Science and Informatics at UCD and a Co-PI on the Strategic Research Cluster in Advanced Geocomputation(StratAG) funded by Science Foundation of Ireland.