Automated Quality Improvement of Road Network in OpenStreetMap

Musfira Jilani  
School of Computer Science and Informatics  
University College Dublin  
Dublin, Ireland  
musfira.jilani@ucdconnect.ie

Padraig Corcoran  
School of Computer Science and Informatics  
University College Dublin  
Dublin, Ireland  
padraig.corcoran@ucd.ie

Michela Bertolotto  
School of Computer Science and Informatics  
University College Dublin  
Dublin, Ireland  
michela.bertolotto@ucd.ie

Abstract

Volunteered Geographic Information (VGI) has recently emerged as a promising technology to create a non-proprietary detailed map of the fast changing world. Since VGI is basically generated by non-professionals the assessment of VGI quality is of utmost importance to determine its suitability for usage. Also there is a pressing need to develop some automated mechanism to improve the quality of this information. OpenStreetMap with a strong and active collaborative community is an active area of VGI research. This paper presents a machine learning based solution for improving OSM quality for road network.

Keywords: VGI Quality, Graphical Representations, Street Network, Machine Learning

1 Introduction

Recent emergence of advanced technologies such as Web 2.0 and Global Positioning Systems (GPS) has given birth to a revolutionary style of mapping. Map making is no more a sole authority of professional cartographers. Today there are six billion human sensors of geographic information. With a mere access to internet and some readily available geo-referencing mechanism anyone can contribute to the generation of geographic information. Goodchild termed this phenomenon as Volunteered Geographic Information [VGI] and asserts that VGI can lead to hitherto unknown, innovative and cheap production and usage of maps [1, 5].

OpenStreetMap(OSM) is by far the most famous example of VGI. Founded by Steve Coast in London in 2004, it is a collaborative project which aims at the creation of free editable map of the world. A rapidly increasing demand for non-proprietary geodata has made OSM highly popular. Several applications based on OSM data such as route planning and geocoding [2], location based services [3] and 3D [4] are being developed. This plethora of crucial applications necessitates the quality assessment and/or improvement of OSM data.

As an outcome of several efforts to study the quality of geographic data, five fundamental dimensions for geospatial data standard have been agreed upon [5]. These include positional accuracy, attribute accuracy, logical consistency, completeness, and lineage. Girres and Touya [6], while assessing the quality of French OSM, also considered semantic accuracy, temporal accuracy and usage as potential data quality indicators.

Several studies have been conducted to study the evolution and quality of OSM. OSM quality has been assessed by comparing OSM with official datasets [6-8]. However, from [9] it can be observed that there can be several local significant features which may be missing from official or proprietary datasets. Also, comparison with official datasets usually involves manual matching of features.

In this paper, a machine learning based solution for improving the quality of OSM data has been presented. In particular the semantic accuracy of road network with special focus on pedestrian and residential roads is being studied. A graph based representation of OSM transportation network has been developed. Artificial Neural Networks (ANN) based framework is being developed to model relationships between data.

The rest of the paper is divided as follows. Section 2 describes the dataset used in the study. Section 3 presents the methodology used for road network quality improvement. In section 4 a discussion on preliminary results is presented. This is followed by conclusions and future work.

2 Dataset

In this paper, residential and pedestrian street networks of OSM corresponding to Dublin and Waterford are studied. Dublin is the capital and most populous city of Ireland with an area of 114.99 km². Waterford is a city located in the South – East of Ireland and has an area of 99.3 km².

The OSM database for Ireland and Northern Ireland was downloaded from Geofabrik (http://www.geofabrik.de/). This data was pre-processed using osm2pgsql tool to obtain OSM data corresponding to Dublin and Waterford. This was further processed by osmfilter to obtain data corresponding to street network only.

3 Methodology

The approach followed is explained by means of a flow diagram shown in Figure 1. A graph based representation
OSM database is drawn from heterogeneous sources and finding patterns in such data is a challenging task. Machine learning has recently provided new insights into the understanding of complex systems like human genome. It is also being used in several systems like self driving cars, speech recognition, email spam, web search, social network analysis etc. All these applications require learning from complex datasets. Hence, it is envisaged that a machine learning approach is a natural way to learn OSM data, determine its quality and suggest an improvement for bad or incomplete data.

Machine learning algorithms are broadly categorized as ‘supervised’ and ‘unsupervised’ depending on how data is being learnt. In the current study, a supervised machine learning approach was used. Assuming that 80 percent of OSM data is good, the training data consisted of residential and pedestrian streets with tag information.

Simple and generic features like degree distribution, betweenness centrality, node count within a bounding box were calculated and are explained as follows. Degree distribution of a graph is a probability distribution of the number of connections of each node over the whole network. Betweenness centrality is a measure of centrality of a vertex within a graph and determines the number of times a node acts as a bridge between two other nodes along the associated shortest path.

Node count within a bounding box was also introduced as a possible feature to support learning. It represents the number of nodes of a graph that fall within a bounding box. All the edges of the graph were traversed sequentially and for every edge the dimensions of corresponding bounding box were calculated dynamically by adding +50 m to the maximum and minimum coordinates present within the edge. The number of nodes of the graph falling within this bounding box were calculated and stored as node count within a bounding box.

Artificial Neural Network (ANN) is a biologically inspired mathematical model developed to imitate human brain’s capability to model data and learn relationships. Fischer and Gopal [10] observed that in situations where data is incomplete, ANNs outperform other statistical models. Therefore, ANNs are expected to perform particularly well with OSM data which suffers from several quality issues like completeness and inconsistency.

A feature vector comprising of the features discussed in previous sections was calculated and is input to an ANN framework which is currently being developed. The ANN model in this study learns the features associated with different types of streets such as residential and pedestrian streets during the training phase. In testing phase ANN detects residential and pedestrian streets without using any tag label. ANN model can now assess the semantic information associated with these streets and suggest improvements.

4 Preliminary Results

An evaluation of the presented approach is currently under development. Initial experiments have shown that betweenness centrality and node count within a bounding box are promising features to use for learning.
Table 1: Average node count within a bounding box statistics for residential and pedestrian streets

<table>
<thead>
<tr>
<th></th>
<th>Dublin</th>
<th>Waterford</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residential</td>
<td>24</td>
<td>30</td>
</tr>
<tr>
<td>Pedestrian</td>
<td>12</td>
<td>10</td>
</tr>
</tbody>
</table>

Residential roads have more street intersections and turns within a small area. Pedestrian roads, on the other hand, are more turn free and hence do not contain large number of nodes in a small area. Experimental results confirm this hypothesis. Table 1 shows that the average number of node count in a bounding box for a residential road far exceeds that of a pedestrian road in both the chosen datasets of Dublin and Waterford. This is pictorially illustrated by means of Figures 2(a) and 2(b).

Pedestrian roads often serve as links of communication between different types of roads and pedestrian roads are more frequently used for travelling as compared to residential roads. The between centrality for pedestrian roads is found to be greater than that of residential roads.

Figure 2(a) : Nodes within a bounding box for residential street

Figure 2(b) : Nodes within a bounding box for pedestrian street

5 Conclusion and Future Work

Topology alone is not sufficient. The current approach doesn’t incorporate context. However, context can provide a very useful learning environment. Also, a cognitive aspect for data modeling can provide some valuable insights to OSM data quality. In future, context and ontologies will also be incorporated. A rule based system will also be developed to aid the machine learning process.

Acknowledgement

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

References


