Automated Highway Tag Assessment of OpenStreetMap Road Networks

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

Padraig Corcoran
Computer Science and Artificial Intelligence Laboratory
MIT, MA, USA
padraigc@mit.edu

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

ABSTRACT

OpenStreetMap (OSM) has been demonstrated to be a valuable source of spatial data in the context of many applications. However concerns still exist regarding the quality of such data and this has limited the proliferation of its use. Consequently much research has been invested in the development of methods for assessing and/or improving the quality of OSM data. However most of these methods require ground-truth data, which, in many cases, may not be available. In this paper we present a novel solution for OSM data quality assessment that does not require ground-truth data. We consider the semantic accuracy of OSM street network data, and in particular, the associated semantic class (road class) information. A machine learning model is proposed that learns the geometrical and topological characteristics of different semantic classes of streets. This model is subsequently used to accurately determine if a street has been assigned a correct/incorrect semantic class.

Categories and Subject Descriptors

E.1 [**Data Structures**]: Graphs and Networks; I.5.4 [**Artificial Intelligence**]: Pattern Recognition—*applications*

General Terms

Reliability, Verification

Keywords

OpenStreetMap, Data Quality, Street Network Analysis, Machine Learning

1. INTRODUCTION

The OpenStreetMap (OSM) project is perhaps one of the most successful examples of crowdsourcing in the spatial domain. Although some parts of the OSM database have been mass imported from sources, such as the TIGER database of the US Census Bureau, most of the data is produced by volunteers with little or no professional skills in effective mapping practices. This raises some genuine concerns regarding the quality of this data and has limited the proliferation of its use.

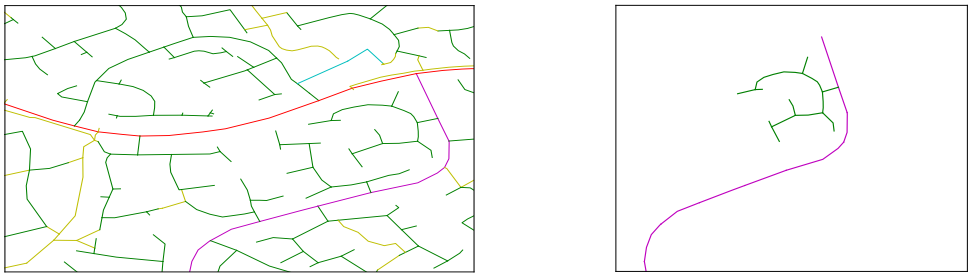
Streets comprise the single most important feature in the OSM database. Street network information is fundamental for many applications such as navigation, network analysis, and map generalization, just to name a few. However, the use of street network data in such applications is only possible if the associated semantic information is correct. The basic semantic information of a street is its class such as a motorway, a primary road, or a residential road, etc. This information indicates several things about a street: its possible neighborhoods, its permissible driving speeds, and the level of map generalization at which it should be displayed. Semantic information in OSM is recorded using tags. An OSM tag consists of a *key=value* pair; here, key represents some feature and value details this feature. In particular, the key 'highway' is commonly used to describe the class of a road (e.g., *highway=residential*). The main objective of the presented study is to assess the quality of the values associated with the highway key in OSM street network data.

Quality assessment is an important research problem: knowing that the data is of high quality allows people to use it with confidence, while knowing that it is of poor quality warns them of potential risks. Most of the current approaches for OSM data quality assessment [5, 6] require referencing to 'accurate' ground truth, which may be unavailable. And even if the ground truth is available, an effective comparison is difficult because both datasets must first be registered, that is, correct correspondence between objects must first be determined before the objects in question can be compared [8]. Moreover, most of the current approaches focus on assessing the geometric accuracy and completeness of OSM data and little has been done to assess its semantic accuracy. In this paper, we propose a fully automated method for assessing the semantic accuracy of the values associated with the highway key of OSM street networks which does not require referencing to ground truth.

In this work we consider connected sets of street segments of uniform semantic class where we define a street segment as the section of a street between two intersections, or between a dead-end and an intersection. We hypothesize that the highway key value for street segments is implicitly rep-

SIGSPATIAL'14, November 04 - 07 2014, Dallas/Fort Worth, TX, USA

http://dx.doi.org/10.1145/2666310.2666476



(a) (b)

Figure 1: A section of street network from OSM London. (a) Five different road classes can be seen: secondary(red), tertiary(magenta), residential(green), footway(yellow), service(cyan). (b) Two connected sets of street segments, one each from tertiary and residential classes have been redrawn for comparisons.

resented in these sets. Figure 1(a) shows a small part of the London OSM street network consisting of several road classes each represented using a different colour. Several simple distinguishing characteristics of these road classes are evident. For example, in Figure 1(b), we compare two connected sets of street segments belonging to two different road classes: tertiary (magenta) and residential (green). The connected set of segments corresponding to the residential class has more dead-ends as compared to the set corresponding to the tertiary class. Similarly, the overall linearity for the tertiary set of street segments is much higher than the residential set where we define linearity as the degree to which the street segments in question have a shape similar to that of a straight line. Furthermore, the residential set is only connected to the tertiary set whereas the tertiary set is connected to both the residential and secondary sets of street segments. This example suggests that the characteristics of connected sets of segments of uniform class vary as a function of class.

The research question we ask is: Can we effectively measure the aforementioned characteristics and subsequently use them to determine street classes? As described above, the characteristics enabling distiction between road classes are a function of a connected set of street segments where all elements of a particular set have the same semantic class. However, such information relating to semantic class is implicitly represented in the street network. Therefore, the measurement of such characteristics requires two steps: firstly, determining the sets which are implicitly represented in the street network and secondly, measuring the characteristics of these sets that allow us to differentiate between different semantic classes. In order to extract connected sets of streets of uniform class we construct a multi-granular representation of the street network [7]. Subsequently we extract characteristics from these sets which allow us to discriminate between different classes. Later, we develop a machine learning model to learn the characteristics specific to various road classes. Finally, we use this model to assess the values associated with the highway tag in the OSM database. Our results show that while it is easy to learn and distinguish between certain classes of roads, there are some overlapping classes (i.e. roads that cannot be distinguished from each other using the proposed characteristics).

2. STREET NETWORK REPRESENTATION

Data representation is often considered as the first step towards knowledge discovery. Street networks are usually represented using a graph. The two predominant graph based street network representations are the primal and the dual forms [2]. We are interested in inferring the road class associated with connected sets of street segments by analysing the geometrical and topological properties of these sets. The primal representation provides such properties but only with respect to single street segments. On the other hand, the dual representation provides us the desired sets but does not represent such properties. Therefore, it is necessary to construct and link both representations. An approach for achieving this was presented by Jilani et al [7] in the form of a multi-granular representation for street networks. This is essentially a two-layered representation which combines both the primal and the dual forms of street network representations. The criterion for set identification of street segments follows a *named-typed* approach: adjacent street segments having same name and class are identified as sets of street segments for the dual graph. At the basic level of granularity, in a multi-granular representation, lies a primal graph. On top of this primal graph rests a dual graph. Adjacent street segments having same name and same road class correspond to a single vertex in the dual graph. Every vertex of the dual graph is a subgraph in itself. All the spatial properties of the entire network are retained. Moreover, the spatial properties for connected sets of street segments become explicit in this representation. While the topological properties of a connected set can be extracted from the top layer (dual), the geometric properties can be obtained by referring to the primal subgraph structure contained within the vertices of the dual graph.

Having obtained the multi-granular representation for a street network, as discussed above, we extract a set of geometrical and topological characteristics for the corresponding sets of street segments that are representative of their semantic class. In the following section we describe the specific characteristics extracted.

3. ROAD CLASS CHARACTERISTICS

The road class characteristics used in this study can be divided into two broad categories: geometrical and topolog-

ical. The choice for these characteristics is based mainly on our observation, domain knowledge, and computational efficiency.

3.1 Geometrical Characteristics

The geometrical characteristics of a road include its shape, size, and other spatial properties such as absolute and relative positions of various road segments. In order to obtain this information for sets of street segments we extracted four characteristic using the multigranular street network representation. These include length, number of dead-ends, number of intersections, and linearity.

Length: Length is a fundamental spatial property and is an important distinguishing characteristic between road classes. Some classes such as residential roads and footways tend to be much shorter in length than others such as motorways, trunk roads etc.

Number of dead-ends: A street with a dead-end or cul-de-sac provides essentially only in/out access but no through traffic. Cul-de-sacs have been studied from urban design and traffic engineering perspectives [3] and can suggest information regarding a street's shape and its role in the overall transportation flow. Certain roads such as residential tend to have more cul-de-sacs as compared to others such as primary and tertiary which carry the vast majority of traffic.

Number of intersections: The number of intersections contained within a set of street segments are an indicative of its shape. The set of street segments with a higher number of intersections is in general more connected.

Linearity: Some road classes such as motorways tend to be more linear as compared to other road classes such as residential. We used the method proposed by Zunic and Rosin [10] for calculating linearity of sets of street segments. This method is especially useful in our case as it does not require strictly defined end-points of the curve in question.

3.2 Topological Characteristics

The most important topological characteristic for a street network is connectivity between various street segments comprising the network. The topological characteristic used in this paper comprises of identifying the semantic classes associated with other sets of street segments directly connected to the set in question. A bag-of-words model [9] for each set of street segments is constructed using the multi-granular street network representation. The words in this bag consist of road classes corresponding to the sets of street segments directly connected to the set in question.

4. LEARNING ROAD CLASSES

Having obtained the characteristics for sets of street segments as described in the previous section, we adopt a data driven approach to learn the characteristics specific to various street classes from OSM data. This is achieved by using a supervised learning methodology to learn from the OSM street network data itself the road class information implicit in the characteristics of the data.

The success of a given supervised learning technique significantly depends on the availability of good training data i.e., data with correct labels. The quality of OSM data for London has been shown to be high [1, 6]. Hence, we have used the London OSM street network data for developing our model. The OSM wiki page lists 42 possible values that can be associated with the highway key. The extracted data contained 38 of these values. 17 of these values were dropped in our experiments owing to the corresponding small proportion of sets of street segments (less than 10 in a dataset consisting of over 70,000 instances). Hence, in the development of our model only 21 highway key values (road classes) have been considered.

The entire London OSM database was divided into three parts: training, validation and test set in 60-20-20 percentage split. Training and validation sets were used for developing the model. The validation set was specifically used to tune the parameters of the model learnt on training data and to ensure model generalization. The test set, as will be described in Section 5, was used for assessing the performance of the developed model on new (unseen) data. Algorithms representative of various machine learning families were compared in terms of ROC area [9] to identify the algorithm that may best tackle the task of learning road classes. The algorithms evaluated were neural networks, support vector machines, naive bayes, nearest neighbour and decision trees. Of these, the decision trees gave optimal performance. We further tried to improve the performance of decision trees by mixing them with ensemble techniques such as boosting and random forest. Of these, the random forest based model performed the best and hence was chosen as the final model for assessing the semantic accuracy of highway key in OSM database.

5. RESULTS AND DISCUSSION

The developed model was applied on two different datasets: firstly, to the test set obtained from the London OSM street network (as described above) and secondly, to the OSM street network of East Sussex (another English city). A demonstration of model performance on a small part of London test set is shown in Figure 2. Correctly classified street segments are represented using solid lines whereas incorrectly classified segments are represented by dashed lines. Classification results for the entire London test set are shown in Table 1. We will now discuss these results in some detail with respect to the classification accuracy and comparison with a baseline.

Accuracy is obtained by dividing the number of correctly predicted instances with the total number of actual instances in the test set for a given class. While the accuracy is quite high (>70 percent) for certain road classes such as motorway links, tertiary links, steps, living streets, unclassified roads, and tertiary roads, the class-wise accuracy for certain road classes such as tracks, secondary links, secondary roads, paths, footways, cycleways, and bridleways can be considered poor (< 40 percent). However, these varying classification accuracies are not surprising. While there are usually 4-10 classes in the overall street network in most of the traditional maps [4], the OSM road classification is quite fine (21 classes considered in this study; 42 available in the OSM catalogue). Hence, some of the road classes in OSM are geometrically and topologically very similar and difficult to model.

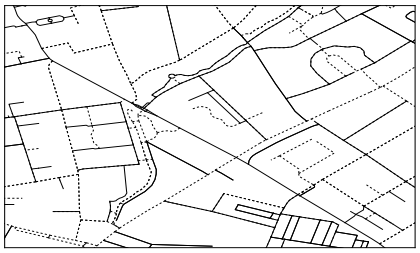


Figure 2: Classification results on a small part of the test set.

Finally, accuracy in itself is not a very meaningful criterion for analysing the classification results as it does not consider the costs associated with misclassification. The complexity of the given problem should be taken into account. Our test data is highly skewed (3 motorways, 8617 residential roads) and we are trying to learn 21 different road classes. Hence, in order to assess the significance of our technique we used a baseline criterion and analysed our model performance with respect to this baseline. We used a frequency baseline which for a given class is the ratio of number of instances contained in this class to the total number of instances in the actual test set. A comparison of the Accuracy and Baseline columns of Table 1 reveals that the developed model performs significantly better than the baseline for all the 21 road classes considered. The results presented here are for the test set obtained from the London OSM street network. Similar results were obtained for the OSM street network of East Sussex.

6. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a machine learning based solution for assessing the semantic quality of OSM data. In particular, we addressed the semantic accuracy of road networks by analysing the values associated with the 'highway' tag. We envisaged that the geometrical and topological characteristics of connected sets of street segments are representative of their semantic class. A novel multigranular street network representation was used to extract these characteristics and a machine learning model was trained to learn them. This model was later applied to test data to determine if the semantic classes associated with streets are correct. Our experiments demonstrated that while it is easy to learn certain street classes such as motorways, motorway links, tertiary roads, tertiary links, steps, living streets and unclassified roads, certain road classes such as secondary roads, secondary links etc cannot be modelled correctly using the proposed charateristics. However, the developed model performs significantly better than a frequency baseline for all 21 street classes considered in this research.

7. ACKNOWLEDGEMENTS

This work was supported by the Irish Research Council through the Embark Postgraduate Scholarship Scheme 2012.

Table 1: Baseline and classification accuracy for the London test set.

Street Class	Baseline	Classification Accuracy
unclassified	8.20	70.60
trunklink	4.00	69.40
trunk	0.30	41.00
track	0.80	24.80
tertiarylink	0.03	100.00
tertiary	2.52	74.90
steps	2.05	83.90
service	17.00	59.80
secondarylink	0.03	16.70
secondary	0.50	20.40
residential	48.00	64.70
primarylink	0.20	59.60
primary	0.99	52.50
pedestrian	1.42	64.20
path	0.90	23.60
motorwaylink	0.05	100.00
motorway	0.02	66.70
livingstreet	0.12	77.30
footway	13.91	34.00
cycleway	1.35	36.40
bridleway	0.17	23.30

8. REFERENCES

[1] A. Ather. A quality analysis of openstreetmap data. *ME Thesis, University College London UK*, 2009.

[2] P. Corcoran, P. Mooney, and M. Bertolotto. Analysing the growth of openstreetmap networks. *Spatial Statistics*, 3:21–32, 2013.

[3] P. Cozens and D. Hillier. The shape of things to come: New urbanism, the grid and the cul-de-sac. *International Planning Studies*, 13(1):51–73, 2008.

[4] G. Forbes. Urban roadway classification: Before the design begins. *Paper presented at the Urban Street Symposium, Texas, Dallas*, TRB Circular E-CO19, 1999.

[5] J. F. Girres and G. Touya. Quality assessment of the french openstreetmap dataset. *Transactions in GIS*, 14(4):435–459, 2010.

[6] M. Haklay. How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and planning*, 37(4):682–703, 2010.

[7] M. Jilani, P. Corcoran, and M. Bertolotto. Multi-granular street network representation towards quality assessment of openstreetmap data. In *Proceedings of the Sixth ACM SIGSPATIAL IWCTS*, 2013.

[8] T. Koukoletsos, M. Haklay, and C. Ellul. Assessing data completeness of vgi through an automated matching procedure for linear data. *Transactions in GIS*, 16(4):477–498, 2012.

[9] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[10] J. Žunić and P. L. Rosin. Measuring linearity of open planar curve segments. *Image and Vision Computing*, 29(12):873–879, 2011.