Mining the Co-existence of POIs in OpenStreetMap for Faulty Entry Detection

Alireza Kashian[1], Kai-Florian Richter[2], Abbas Rajabifard[1], Yiqun Chen[1]

[1] Department of Infrastructure Engineering, University of Melbourne, VIC, Australia, Emails: a.kashian@student.unimelb.edu.au (A.K.); abbas.r@unimelb.edu.au (A.R.); yiqun.c@unimelb.edu.au (Y.C.)
[2] Department of Geography, University of Zurich, Zurich, Switzerland, Email: kai-florian.richter@geo.uzh.ch (K.R.)

SUMMARY

In recent years, more and more volunteers join crowdsourcing activities for collecting geodata which in turn might result in higher rates of man-made mistakes in open geo-spatial databases such as OpenStreetMap (OSM). While there are some methods for monitoring the accuracy and consistency of the created data, there is still a lack of advanced systems to automatically discover misplaced objects on the map. One feature type which is contributed daily to OSM is Point of Interest. In order to understand how likely it is that a newly added POI represents a genuine real-world feature, some means to calculate a probability of a POI existing at that specific position is needed. This paper reports on work in progress on a platform for analysing POI objects in the OSM database in order to find patterns of co-existence among features in close distance to each other. These patterns will improve current tracking and verifying systems and, thus, enhance positional accuracy of registered POIs in OSM.

Keywords: OpenStreetMap, VGI, POI, Geographic Information Quality, Feature Selection, Pattern Mining, Correlation, Tobler's First Law

INTRODUCTION

Advances in positioning, web mapping and communication technologies have given us better leverage to tackle the incompleteness of spatial data. People distributed within an environment can voluntarily participate in collaborative online activities and help producing knowledge about every object which has geographic attributes. Goodchild [1] coined this phenomenon VGI (Volunteered Geographic Information) which encapsulates the idea of collecting, maintaining, and distributing geographic information with the help of volunteers.

But the increasing popularity of VGI activities also comes with a number of caveats. Making mistakes or malicious activities have a long history in crowdsourcing environments [2-4]. While statistical reports indicate that most contributions in VGI projects are genuine, incorrect or inaccurate data entries and even suspicious activities by spammers in planned organized attacks are on the rise [5].

One major category of such malicious activities or non-deliberate mistakes in VGI projects is the creation of non-existing features or the modification of positions and attributes of objects, which potentially impacts the quality of the generated data and consequently reduces the credibility and viability of such systems. Our work presents an innovative research tool to analyse and discover meaningful patterns among recorded objects in OpenStreetMap (OSM) with a specific focus on Points of Interest (POIs). This allows us to identify untrustworthy data entries and in turn improves the reliability of VGI platforms by making them sufficiently intelligent in providing credible information used for serving a variety of location-based applications in the future. In this short paper, we briefly review the OpenStreetMap project and some of the common faults made by amateur volunteers and then introduce the statistical pattern discovery tool which processes millions of records to extract co-existence rules among POIs. Finally, we present an outlook on the results achievable with our tool.

SPATIAL DATA QUALITY AND CO-EXISTENCE OF POIS

POIs have a central role in most of today's online mapping applications and offline navigators. They also play a critical role in location-based social networking applications, such as Foursquare and Yelp, where people look for tips and reviews written by other peers. Incorrect position or inaccurate

attribute information for such locations will discourage users to continue using these services as the information ambiguity will practically reduce the level of trust. Quality assessment of geographic data generated in VGI projects has been the focus of research in the past decade [6-9]. Spatial data quality comprises several basic elements; among them logical consistency and positional accuracy play an important role to guarantee the quality of data. While some researchers have tried to demonstrate that completeness of geodata in OSM is comparable to authoritative sources by measuring the gaps in two overlapping datasets [10], others recently have paid attention to logical inconsistencies in OSM by developing a framework based on the concept of spatial similarity in three dimensions, i.e. directional relationships, topological relationships and metric distance relationships [11].

In this work, we focus on positional accuracy of POIs in OSM. We emphasize Tobler's first law of geography, which claims that everything is related to everything else but nearby things are more related than distant things [12]. Based on this law, we aim to discover potential co-existence patterns among POIs and other geographical features, such as roads and buildings, which are in close proximity to each other. For example, consider the relationship between gas stations and road segments. As we all know vehicles need access to road structure to drive into gas stations. We would assume that whenever we find a gas station, it is highly likely to have a road segment nearby as well. Classical data mining algorithms [13] are often based on assumptions which violate Tobler's law (e.g. independent, identical distributions). Nearby objects in a spatial context tend to affect each other rather than acting independently.

The development of robust and innovative tools to extract useful information from existing geo-spatial datasets is crucial for any organisation that has to make critical decisions based on large spatial data sets. This also holds for large crowdsourcing datasets, such as OSM, where a quality control mechanism is recommended to monitor the contributions at early stages of data creation or data modification instead of piling up hundreds of thousands of unverified records into existing databases. Fortunately, most VGI platforms have already incorporated different control mechanisms within their editors for quality assurance and validation. In the case of OSM, several free online quality assessment and assurance tools have been developed to get detailed quality information. Interested users are able to report errors in the data by using OSM Notes or OpenStreetBugs. Other tools, such as Keep Right, Osmose or OSM Inspector, can be used to visualise detected errors in the map data. The JOSM editor informs a user prior to the upload if there are any intersecting geometries or duplicated elements. However, these editors only inform the user, but do not refuse to actually upload the changes [5].

And all these mechanisms are based on geometry only. What is missing is a 'semantic' analysis of newly contributed data. Hence, we study co-existence patterns for POIs with two potential aims in mind: 1) to help the OSM community with a new advanced monitoring tool to identify mis-located POIs and to highlight them for volunteer editors for correction; 2) to help urban planners to discover which objects are inconsistently located in the city and plan for a better future distribution of service locations and goods delivery.

OPENSTREETMAP

In our work, we use the OpenStreetMap database to analyse and extract patterns for finding the co-existence relationship between two different POI types in a given city. The OpenStreetMap (OSM) project started in 2004 with "building a global map" as the main aim of the project [14]. More than 2 million users had joined this project by the end of 2015. Any registered member can add and edit geographic objects without any restrictions. This method of data collection is in some contrast with other VGI projects, such as Google Map Maker, where the alterations made by new members are reviewed first before being applied to Google Maps. But OSM provides open access to all recorded data. Additionally, it provides access to historical changes for each individual object so any unwanted changes can be rolled back easily by supervisors.

Before talking about potential faulty entries in the OSM project, we briefly introduce the four fundamental data elements in OSM, which are[*]:

Node: A node represents a specific point on the earth's surface defined by its latitude and longitude. Each node comprises at least an id number and a pair of coordinates. Nodes can be used to define standalone point features. Nodes are also used to define the shape of a Way.

[*] http://wiki.openstreetmap.org/wiki/Elements

Way: A way is an ordered list of between 2 and 2,000 nodes that define a polyline. Ways are used to represent linear features, such as rivers and roads. Ways can also represent the boundaries of areas (solid polygons), such as buildings or forests.

Relation: A relation is a multi-purpose data structure that documents a relationship between two or more data elements (nodes, ways, and/or other relations). An example is a turn restriction at an intersection.

Tag: All types of data element (nodes, ways and relations) can have tags. Tags provide meaning (the semantics) for a particular element to which they are attached. A tag consists of two free-form text fields; a 'key' and a 'value'. For example, "highway=residential" defines the way as a road whose main function is to give access to people's homes.

TYPES OF FAULTS IN OSM

There are some common mistakes, which are often made by OSM volunteers. Some of these faults are considered as vandalism, which corresponds to malicious activities of spammers, while the rest just emerges from wrong assumptions or wrong interpretation of satellite imageries while inserting objects on the map. Distinguishing abnormal edits, such as the addition of new a POI at a wrong location, and highlighting them for further inspections is beneficial to human editors. The OSM Wiki lists the following potential error sources:

- Addition of a new object with has new attributes with no previous history in the database
- A modification of geometric attributes of an object in a non-regular format
- A non-routine change in the attributes of an object
- Bulk removal of existing features
- An abnormal behaviour, such as editing specific attributes only by a single user
- Inserting non-existing features on the map
- Using scripts or other bots to do automatic bulk edits inappropriately
- Doing script-like edits (e.g., selecting 10 parks and inserting the key park: tree: type = none)

OSM POI ANALYSER PLATFORM

We designed and implemented a new framework to analyse the currently existing POIs in OSM. The main function of the platform is to evaluate if the position of a new POI is likely acceptable. We devised a mechanism for measuring the probability of a particular type of POI existing at the proposed position, which is measured against all other existing similar objects in the same city. For example, logically we expect a carwash service to be always close to a road segment, or a ferry terminal to be somewhere very close to a body of the water, such as a lake or a river. This kind of knowledge can be systematically generated by mining all ferry terminals or all carwash services and then establishing their relationships with nearby objects, such as roads, rivers and lakes. In other words, access to interpretable and meaningful knowledge about our existing world is critical for finding meaningful relational patterns. We are interested in knowing how cities are organized, specifically for public service access points, where specific geographic features are usually located, which pair of objects is close to each other most of the times, and even discover whether some objects are dependent on the availability of some other objects. Among these questions, we focus on the co-existence of pairs of objects. Due to differences in the development of cities and city planning, we expect to see different association rules between objects in each individual city, so we would expect most rules to be valid only for particular cities, while there might still be others that can be applied globally.

To establish these rules, a comprehensive analytic platform was designed and developed using the latest data from the OSM database. The platform was implemented in PHP using a PostgreSQL database on a Debian cloud server. For the pilot tests we focus on those cities that had the highest editing activity (for 30 days in May 2015). This resulted in Paris, Madrid, Toronto, Frankfurt and Warsaw as our test cases.

THE PLATFORM

Our platform, called OSM POI Analyser, processes data using 15 nearest neighbourhood ring regions. An overview of the platform is shown in Figure 1. The platform is accessible at http://validate.openstreetmap.me

We use spatial clustering in order to classify nearby objects into 15 classes with known distance ranges. As an example, objects within 100 to 200 meters range form one distance class in our analysis. The 15 circular regions are further illustrated in Figure 2. These 15 distance regions reflect Tobler's First Law, which states that nearby objects have a stronger relationship with each other. If an ATM

machine is 5km away from a bank office, then we cannot expect to see any significant relation between this pair of objects, but if most ATMs in a city were within 50 meters distance of a bank, then we would say that a potential association is observable here.

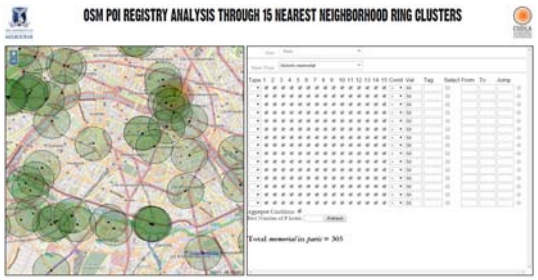

Figure 1. Main screen of the OSM Analyser platform.

Figure 2. 15 neighbourhood ring regions. 10 m range for the first 5 rings and then 100 m for the rest.

The distance regions, which form doughnut rings, have different ranges. The first 5 doughnut regions have a 10 meters range, the 6th ring has a 50 meters range, and all the rest have a 100 meters range. Since the 15th region ranges from 900 meters to 1000 meters, objects beyond 1,000 meters are ignored in our processing. The 1,000 meters distance was selected as a cut off point for reducing processing of unnecessary objects, which likely would have no or almost no correlation with the POI under consideration. Future work will need to tell whether we miss important relationships using this cut-off point. Processing all POIs of a given type that exist within a city we will be able to answer questions such as: What percentage of ATM machines is within 50 meters distance of banks in Paris? What percentage of ferry terminals is within 100 meters distance to rivers in London?

We used the platform to pre-process data from the five selected pilot cities. Prior to processing, the following data preparation procedure was performed:

- A list of known tags for POIs in OSM was prepared. Some tags are officially listed by the OSM project as commonly accepted tags for specific features on the map. This selection of tags helped us identifying the nodes which actually represent POIs as there are many nodes, which are only part of a way and do not represent a POI.
- The complete datasets for the sample cities were imported from OSM. This includes Nodes, Ways, Tags and Relations.
- All POIs were extracted in separated tables for each city.
- To reduce the amount of computation, only a sample subset of POI types were selected for further processing. The sample subset includes 22 manually selected types, such as amenity:ATM, amenity:bank, office:company, leisure:playground, and amenity:post_box. The selection criteria was based on those object types which were frequent and we expected to find them normally throughout the city as it was enabling us to consider all possible patterns in different locations. So object types such as airports are not selected.
- For each POI type (e.g. amenity:ATM), all instances were extracted for each city. For example, there are 3,944 Traffic:Signals in Madrid or 1,890 Amenity:Bank in Paris.
- For each individual instance (single POI), we created the 15 circular ring regions around that instance.
- Queries were performed to extract any object (Way, Node), which is within or intersects with each ring region. The results were recorded in a new table for each city and showed which objects are located around the target instance. For example, for each bakery in Paris, an intersection query for each ring cluster is performed. If Paris has 2000 bakeries, it means that intersect queries are run 2,000 x 15 = 30,000 times.

Using our platform, a user first picks one of the five pilot cities, then selects a POI type from a drop down menu, and finally clicks on a location on the map to register a new POI of that type. By clicking on the map the user proposes this location for the new POI. Next, our system runs a query and extracts all processed information about this POI type from the recorded data and, concurrently, it evaluates the newly proposed location by checking all nearby objects within the 15 ring clusters. All processed information is reported in different tables on the screen. In more detail, the platform performs the following steps for a new POI registration:

Step 1: Analysis of nearby objects to find if duplicate object(s) exists. The user will be informed if duplicates are found. Duplicates get reported if a similar POI type is observed within less than a few meters distance. This duplication distance was established by processing hundreds of POI instances in

the database to discover the minimum, maximum and average distance between two POIs of the same type. For example, the minimum distance between two ATM machines in Paris is only 1 meter and the average is 240 meters.

Step 2: Checking for overlap with the buffer of nearby objects. For example, a highway has a 20 meters buffer from its centre and the system avoids registration of a hospital, which intersects with this buffer. The buffer size is different for each object type, which was manually set in a configuration table.

Step 3: Checking whether the pattern of relationships with objects in the distance rings is similar to the pattern emerging from the pre-processed data.

To simplify interpretation of the correlations between a POI and all other nearby objects, we divided all geographic features into two separate sets. The first set includes all geographic features, which are mainly used for our daily tasks or we might live, work or shop in those locations. Hospitals, houses, bus stops, monuments and police stations are examples of such instances. The second set covers all other features, which we use as means of transportation or we simply pass through them to get from some point A to some point B. Roads, rivers, lakes and forests are examples of this second set. We term the first set RT (Referring To) and the second set GT (Going Through). There is no feature that belongs to both RT and GT, so the intersection of the two sets is empty. With these sets we are able to extract meaningful relations between both sets. For example, we can see that most of the times gas stations (RT) are close to a road segment (GT), or that ferry terminals (RT) are close to rivers (GT). Aside from relationships between RT and GT sets, there are also interesting relationships between members of the RT set. For example, most of the times an emergency clinic (RT) is close to an existing hospital (RT), or ATM machines (RT) are close to bank branches (RT). These are only some examples, and there are more interesting patterns yet to be discovered.

RESULTS OF A CASE STUDY

As a first case study, we decided to register a new bank in Paris. In Figure 3 the user clicked on some position on the map and the associated clusters are shown around that position. Figure 4 shows a sample co-existence pattern between residential roads and all 1890 banks in Paris. Reviewing the chart, we discover that 90% of banks in Paris have at least one residential road that is found between 300-1000 meters away. The chart also shows that with 55% support, a residential road is found in 40 to 50 meters distance of a bank, which means one out of two banks in Paris is located close to a residential road.

Figure 3. Proposed bank position in Paris.

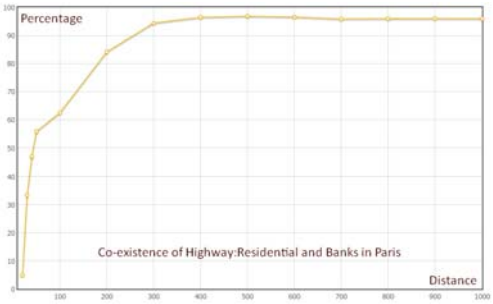

Figure 4. Co-existence of residential roads and banks in Paris.

The platform also reports if a bank already existed near the chosen position, or which objects are found in 20 meters or 500 meters distance. We can also identify which objects around all 1890 bank branches in Paris have the highest average support. For instance, as seen in Table 1, 62.45% of the time (on average), a highway:crossing exists within 1000 meters distance of a bank in Paris. The platform can also generate more detailed reports, such as:

- If duplicate objects are found,
- Objects that are very close to the newly registered point (within its buffer size),
- Objects that exist around the current proposed point and were seen before with other objects of this type,
- Objects that do not exist around the current proposed point but were seen before,
- Objects that exist around the current proposed point but were not seen before,
- Objects inside which the registered point is located in (geometrically),

- Which objects are located in each of the distance rings (all objects, or only GT or RT objects),
- Average item count for Going Through objects (e.g., roads, lakes, rivers, forest),
- Average item count for Referring To objects (e.g., hospitals, bus stops, sport complexes, churches).

With this information at hand, many more questions could be explored, such as:

- Do the banks in all five cities have similar co-existence patterns with other surrounding features?
- Which features always exist within 1000 meters distance of banks in Paris?
- What other POIs are mostly observed around banks in Paris?
- Which banks in Paris do not have a building around them?

Tag:Value	Support Percentage (1 km range)
Building:yes	98.29%
Landuse:residential	82.98%
Highway:crossing	62.45%
Highway:bus_stop	49.64%
Highway:traffic_signal	40.60%

Table 1. Top 5 features with highest average support around banks in Paris.

FUTURE WORK

So far, we do not take into account that the existing OSM data may already be erroneous. To address this, we are planning to test our method using different random subsets of POIs and to compare the extracted patterns regarding their robustness. We are also planning to extend our system such that it traces the registration of POIs in (more or less) real time and, thus, may raise warnings online for human editors, particularly those tracing OSM data quality.

ACKNOWLEDGMENTS

We appreciate the support and valuable comments received from members of the Centre for Spatial Data Infrastructure and Land Administration as well as the Centre for Disaster Management and Public Safety at the University of Melbourne to support our research.

REFERENCES

1. Goodchild, M.F., Citizens as sensors: the world of volunteered geography. GeoJournal, 2007. 69: p. 211-221.
2. Javanmardi, S., et al. User contribution and trust in Wikipedia. in Proceedings of the 5th International ICST Conference on Collaborative Computing: Networking, Applications, Worksharing. 2009. IEEE.
3. Mola-Velasco, S.M. Wikipedia vandalism detection. in Proceedings of the 20th international conference companion on World wide web. 2011. New York, NY, USA.
4. Tang, X., et al., Detecting Wikipedia Vandalism with a Contributing Efficiency-Based Approach., in Web Information Systems Engineering-WISE 2012. 2012, Springer Berlin Heidelberg. p. 645-651.
5. Neis, P., M. Goetz, and A. Zipf, Towards Automatic Vandalism Detection in OpenStreetMap. ISPRS International Journal of Geo-Information, 2012. 1: p. 315-332.
6. Fan, H., et al., Quality assessment for building footprints data on OpenStreetMap. International Journal of Geographical Information Science, 2014. 00: p. 1-20.
7. Ahmed Loai Ali, F.S., Data Quality Assurance for Volunteered Geographic Information, in Geographic Information Science, E.P. Matt Duckham, Kathleen Stewart, Andrew U. Frank, Editor. 2014, Springer International Publishing: Switzerland. p. 126-141.
8. Goodchild, M.F. and L. Li, Assuring the quality of volunteered geographic information. Spatial Statistics, 2012. 1: p. 110-120.
9. Stark, H.-j., Quality assessment of volunteered geographic information using open Web map services within OpenAddresses. GI Forum, 2011: p. 101-110.
10. Haklay, M., How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and Planning B: Planning and Design, 2010. 37: p. 682-703.
11. Hashemi, P. and R.A. Abbaspour, Assessment of Logical Consistency in OpenStreetMap Based on the Spatial Similarity Concept, in OpenStreetMap in GIScience. 2015, Springer. p. 19-36.
12. Tobler, W., Cellular geography, in Philosophy in geography. 1979, Springer. p. 379-386.
13. Agrawal, R. Tutorial database mining. in Proceedings of the thirteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. 1994. ACM.
14. Data, S., et al., Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. Annals of the Association of American Geographers, 2012. 102: p. 571-590.