Annotating Spatial Features in OpenStreetMap

Peter Mooney and Padraig Corcoran

Department of Computer Science, National University of Ireland Maynooth,
Maynooth,Co. Kildare. Ireland.
email {peter.mooney,padraig.corcoran}@nuim.ie
Tel: 353 (1) 2680100, Fax: 353 (1) 2680199
http://www.cs.nuim.ie/~pmooney

ABSTRACT: OpenStreetMap (OSM) is, potentially, the most famous example of Volunteered Geographic Information (VGI) on the Internet today. OSM volunteers contribute spatial content to the global OSM database. These contributors are encouraged to 'tag' content under the guidance of a flexible community endorsed ontology of spatial object tags provided on the OSM Wiki. This paper explores how OSM contibutors edit spatial objects with an emphasis on 'tagging' generated from an examination of the historical evolution of geographic features such as lakes, rivers, roads, forests, in OSM.

KEYWORDS: OpenStreetMap, Quality, Web GIS, VGI

1. Introduction

The OpenStreetMap object (id $= 24015216$ - see Forest (2010)) is a forest near Stuttgart in Germany. It was first contributed in April 2008. Currently (December 2010) it has recieved 43 subsequent revisions the last of which was performed in August 2010. Upon closer inspection of the history of the revisions to this feature one sees that the object was repeatedly tagged as either "landuse=forest" or as "natural=wood" and on one occasion as a "highway=primary". Depending on when other users downloaded the OSM data for this region, for mapping and/or GIS analysis, they would have been presented with a different tag for this object. Which version is correct? The ambiguity between versions is one example of data quality issues surrounding the OpenStreetMap (OSM) project. OSM is a collaborative project to create a fully free and openly accessible map of the world. Volunteers in the OSM community collect geographic information and submit this to the global OSM database (Ciepluch et al.; 2009). Currently OSM is used primarily for rendering various map visualizations (Auer et al.; 2009). Over et al. (2010) comment that the greatest obstacle to more wider use of OSM is spatial data inhomogenity which is preventing OSM being used in "geomatics applications". Contributors are the cornerstone of OSM. This paper provides result of an analysis of contributions to two large OSM databases.

2. Working with OpenStreetMap data

OSM data is represented by adhering to a relatively simple data model comprised of three basic elements - nodes, ways and relations (Auer et al.; 2009). The creator of OpenStreetMap Steve Coast comments that "metadata in OSM is open ended and simple" (Coast; 2010). There are no restrictions whatsoever regarding the use of attributes or attribute values to annotate elements in OSM. Consequently users can create arbitrary attributes or attribute values (Auer et al.; 2009). The Map Features page on the OpenStreetMap Wiki

(OSM; 2010) describes the OSM community agreed ontology of terms, or "Tags", to describe the geographical features in the OSM database. However the use of the ontology is not strict and is provided as a guide. Coast (2010) comments that by not constraining contributors with an ontology two things are possible. First, creative and unexpected types of geo data can be added and secondly one exposes a playful aspect of the project which is to allow experimentation. This is broadly in agreement with Wang et al. (2010) who remark that "people seem to like these collaborative projects because they enjoy the openness of social media". In relation to the example presented at the begining of this paper Comber et al. (2005) comment that as "the number of nonspecialist users of GI increases and spatial data are used to answer more questions about the environment, the need for users to understand the wider meaning of the data concepts becomes more urgent". Some authors have found that "indexing content through tagging is prone to unsystematic and inconsistent metadata that can potentially harm retrieval performance and make further analysis difficult" (Wang et al.; 2010). In this paper we provide some results from an analysis of tagging and editing of geographic features in OpenStreetMap databases by analysing the history of edits to over 4000 geographic objects. There is little literature published, to our current knowledge, which deals specifically with the metadata or tagging behaviour of contributors to OpenStreetMap. A considerable body of literature exists on tagging and annotation behaviour of contributors in social networks and enterprise systems such as Flickr, YouTube, Delicious, etc. This literature provides very helpful and informative support studies. The OpenStreetMap ontology is best described as a *folksonomy* where "folksonomies and social tagging" provides a cheaper and more natural way of organising web objects (Gupta et al.; 2010). In Gupta et al. (2010) the authors describe a folksonomy as (folk (people) + taxis (classification) + nomos (management)) - a user-generated classification emerging from a bottom up consensus. Crucially unlike formal taxonomies the concept of a folksonomy is one where no explicity relation is defined between terms. It has been shown that since users themelves tag the objects, sometimes from a suggested list of possible terms, the folksonomy directly reflects the vocabulary of the user/contributor (Milicevic et al.; 2010). Ambiguity arises in folksonomies because different users apply different tags to different objects. Acronyms can lead to ambiguity as can spelling errors or the combination of several words into a single word tag. In OpenStreetMap objects are permitted multiple tags. However each tag can only be assigned a single value. In the next section we analyse how annotation and edits of spatial objects in OSM are performed.

3. Experimental Analysis

OSM data is freely available, in OpenStreetMap XML format, from the GeoFabrik website `http://download.geofabrik.de/`. This data is updated almost hourly so the most up-to-date version of the OpenStreetMap database is always available. We downloaded the OSM-XML for England and Germany's Baden-Württemberg region. These two locations were chosen because they have two of the most active OSM communities in Europe subsequently providing us with a large set of geographic objects for analysis. This could be easily extended to other regions with less active OSM communities. We extracted all `ways` with at least 20 versions of edits. For England there are 3250 ways with at least 20 versions while Baden-Württemberg has 909 such ways. For each way w we compute a number of characteristics for each version wv of w including: wv_n - the number of nodes in wv, wv_T - the set of tags (key,value) pairs annotating wv, wv_u - the user id of the user who created wv, and wv_t the timestamp of the edit of wv. Table 1 shows the distribution of time between consecutive edits wv_{t1} and wv_{t2} for all ways in both OSM datasets. There are a number of interesting observations. Almost 42% of consecutive edits are separated by an editing time of 1 week to 1 month. Almost 38% of consecutive edits have 1 hour and 24 hours between them.

There are a number of additional observations, extracted from this analysis, worthy of further discussion. These are summarised as follows:

- In Baden-Württemberg 93 (10%) of polygons (from 909) have no tags at some stage of their evolution

Table 1: Distribution of time between consecutive edits to all ways

No. Edits	**% of Edits**	**Time Between Edits**
3866	3.30%	$\leq$ 5 minutes
11478	9.80%	5mins $\leq$ 30mins
2400	2.05%	30mins $\leq$ 1hour
12183	10.40%	1hr $\leq$ 2hr
21318	18.20%	2hr $\leq$ 12hr
11608	9.91%	12hr $\leq$ 24hr
3391	2.90%	24hr $\leq$ 1week
49084	41.91%	1week $\leq$ 1month
1784	1.52%	$>$ 1month

- In England 531 (16%) of polygons (from 3250) have no tags at some stage of their evolution
- The name of a geographic feature is one of the most basic metadata attribute values for spatial data. 2827 objects use the `name` tag. The number of objects tagged with the `name` tag at their first version is 1190 or (42%)
- Of the objects which use the `name` tag 823 of these objects are first tagged with a name tag as late as the 10^{th} version of their evolution
- In total 114 objects are created with no `name` tag but their current version contains a valid value for the `name` tag
- There are 3332 unique editors. The mean number of editors for all objects is 5.892. The median number of editors of 5.00 with a standard deviation of 3.794. This goes some way towards supporting the anecdotal claim that only a small number of editors do most of the editing in OSM.
- Only 30 objects have twenty or more editors - 7 of these objects represent road polylines with the remainder representing forest features
- There are 120929 unique object versions. 231 users contributed 71% of these versions. 39 "super contributors" (those who contributed more than 500 edits) are responsible for 39.5% of all edits. Haklay et al. (2010) argue that in OSM there is a "decreased gain in terms of positional accuracy when the number of contributors passes about 10 or so".
- In table 2 an example is provided of where the "name" tag of a street (located in Southend-on-Sea, England) changes multiple times. The current version is $v26$. The name changes 4 times from it's original "Thames Drive" to current version of "Grande Parade". Three distinct users are involved.
- Changes to Name Tags: There are 285 polygons in our test set which exhibit a "name" tag which is changed 3 or more times. For each way we clustered the assigned name tags into chronological groups and then compared the transformation of tags into one another using two well known string matching metrics to quantify how similiar the name tags were. The Levenshtein distance is defined as the minimal number of characters you have to replace, insert or delete to transform from one string to another (Yujian and Bo; 2007). The JaroWinkler distance (Bilenko et al.; 2003) is a similiar metric used mostly for duplicate detection in databases. The metric is normalized such that 0 equates to no similarity and 1 is an exact match between the two strings. In Figure 1 we show a plot of the mean

Table 2: An example of the “name” tag changing on a street polyline (4803031) in OSM

Version	Name Tag	User ID	Date of Edit
v1	NULL	64941	23/06/2007
v3	Thames Drive	64941	06/09/2008
v5	Belton Way West	64941	06/09/2008
v21	Belton Gardens	20573	28/02/2009
v22	New Road	20573	28/02/2009
v26	Grande Parade	320358	24/07/2010

Levenshtein distance against the mean JaroWinkler distance for each qualifying spatial object. Most objects are clustered around a mean Levenshtein distance of 10 and mean JaroWinkler distance of 0.5 which indicates that the changes from one name tag to the next name tag are substantially different. This is potentially caused by contributors : spelling placenames incorrectly, providing local variations on official placenames, incorrect naming of streets, and correction or spelling.

- The polylines representing “highways” were analysed. In our test set there are 2889 polylines tagged as highways (trunk,motorway,residential,etc). Of these highway polygons 1143 changed highway designation at least once - for example their tag changed from primary to secondary. Close inspection of these 1143 polylines show interesting tagging behaviour: 594 changed designation once, 293 changed twice, 127 changed three times, 60 changed four times. The remaining 69 polylines have between 5 and 10 designation changes. Incredibly three polylines exist with 23, 41, and 73 designation changes.

- There are 548 polygons tagged as “landuse” polygons. There is less tag changing amongst these polygons. Only 40 of these polygons experience changes to their original landuse tag.

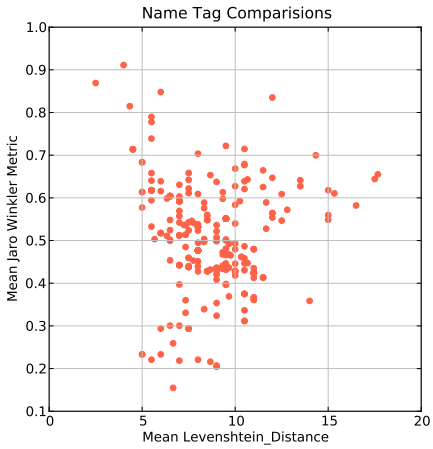

Figure 1: Using the Levenshtein distance and JaroWinkler distance metrics to visualise changes to name tags of spatial objects in OpenStreetMap

4. Conclusions and Further Work

This paper has described the results of an analysis of how spatial features are edited and annotated in OpenStreetMap by extracting the entire history of all contributions to spatial features in two large OSM databases. The majority of the quality analysis of OSM reported (such as Haklay (2010); Mooney et al. (2010)) in the literature base their analysis on the *current* available version of the OSM database or a recently downloaded version of the database. For users the problems associated with annotation of spatial features in OpenStreetMap are of great importance. Potential users of OSM data will require some measures of certainty that the current version of the OSM data has *evolved* to a stable and agreed-upon representation of the real-world features modelled in the data. The changing of names of features (by different contributors), for example, could cause these users to lose confidence with OSM. An example of this is shown in table 2. Section 3. showed that these OSM databases are continuously evolving and changing. The spatial characteristics and the attribute metadata can change quickly often within a very short period of time. While the results of only two regions are presented in this paper the analysis is not restricted to these regions. Any OSM database can be provided as input. The computational-based analysis of how tags describing names or highway designations change is very informative and shows in the case of some features that there is diagreement and ambiguity surrounding the naming of features from local contributors. Local knowledge or OSM contributor input to *why* names changed is needed more completely understand our initial observations made. Future work will pursue a number of research questions including: an analysis of how different users change and edit spatial attributes for certain objects; the correlation between the number of contributors and the number of changes (locally and globally); and finally to find and subsequently quantify evidence of "tag wars" where contributors constantly disagree about the correct values for tags for certain objects.

5. Bibliography

Auer, S., Lehmann, J. and Hellmann, S. (2009). Linkedgeodata: Adding a spatial dimension to the web of data, *in* A. Bernstein, D. R. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta and K. Thirunarayan (eds), *International Semantic Web Conference*, Vol. 5823 of *Lecture Notes in Computer Science*, Springer, pp. 731–746.

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P. and Fienberg, S. (2003). Adaptive name matching in information integration, *Intelligent Systems, IEEE* **18**(5): 16 – 23.

Ciepluch, B., Mooney, P., Jacob, R. and Winstanley, A. C. (2009). Using openstreetmap to deliver location-based environmental information in ireland, *SIGSPATIAL Special* **1**: 17–22.

Coast, S. (2010). Openstreetmap: The best map?, OpenGeoData.org - `http://opengeodata.org/openstreetmap-the-best-map` Last Checked - December 2010.

Comber, A., Fisher, P. and Wadsworth, R. (2005). What is land cover?, *Environment and Planning B: Planning and Design* **32**(1): 199–209.

Forest (2010). A forest feature in the openstreetmap database - located in northern germany, Spatial Content in the OpenStreetMap database - `http://www.openstreetmap.org/browse/way/24015216`.

Gupta, M., Li, R., Yin, Z. and Han, J. (2010). Survey on social tagging techniques, *SIGKDD Explor. Newsl.* **12**: 58–72.

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**(4): 682–703.

Haklay, M., Ather, A. and Basiouka, S. (2010). How many volunteers does it take to map an area well?, *in* M. Haklay, J. Morely and H. Rahemtulla (eds), *Proceedings of the GIS Research UK 18th Annual Conference*, University College London, London, England, pp. 193–196.

Milicevic, A. K., Nanopoulos, A. and Ivanovic, M. (2010). Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions, *Artif. Intell. Rev.* **33**: 187–209.

Mooney, P., Corcoran, P. and Winstanley, A. C. (2010). Towards quality metrics for openstreetmap, *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, ACM, New York, NY, USA, pp. 514–517.

OSM (2010). The map features page on openstreetmap.org, Online Wiki Page: `http://wiki.openstreetmap.org/wiki/Map_Features` (checked December 2010).

Over, M., Schilling, A., Neubauer, S. and Zipf, A. (2010). Generating web-based 3d city models from openstreetmap: The current situation in germany, *Computers, Environment and Urban Systems* **34**(6): 496 – 507. GeoVisualization and the Digital City - Special issue of the International Cartographic Association Commission on GeoVisualization.

Wang, J., Clements, M., Yang, J., de Vries, A. P. and Reinders, M. J. (2010). Personalization of tagging systems, *Information Processing & Management* **46**(1): 58–70.

Yujian, L. and Bo, L. (2007). A normalized levenshtein distance metric, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **29**(6): 1091 –1095.

6. Biography

Dr. Peter Mooney is a research fellow at the Department of Computer Science NUI Maynooth and he is funded by the Irish Environmental Protection Agency STRIVE programme (grant 2008-FS-DM-14-S4). Dr. Padraig Corcoran is a lecturer and post-doctoral researcher also at the Department of Computer Science NUI Maynooth. Dr. Corcoran is part of STRAT-AG which is a Strategic Research Cluster grant (07/SRC/I1168) funded by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.