Similarity Comparisons for Spatial Regions in Volunteered Geographic Information

Peter Mooney*[1] and Nikola Davidovic †[2]

[1]Department of Computer Science, Maynooth University, W23 F2H6 Maynooth, Ireland
[2]Faculty of Electronic Engineering, University of Nis, 18000 Nis, Serbia

Summary

While there is an ever increasing amount of research attention into Volunteered Geographic Information (VGI) there has been little work carried out into comparing different regions or areas in VGI. Previous work has concentrated on comparing VGI with other sources of spatial data. Given two regions A and B from a given VGI project how can we use the annotation of objects in these areas to compare these regions to measure or understand their similarity? In this paper we explore some potential approaches to addressing this difficult problem by applying four metrics to case-study data from OpenStreetMap.

KEYWORDS: Volunteered Geographic Information, similarity, comparison, data quality.

1 Introduction

Over the past number of years there have been many efforts to develop useful comparisons between sources of Volunteered Geographic Information (VGI) and authoritative sources of geospatial data such as that from commercial mapping companies, National Mapping and Cadastral Agencies or government databases (Olteanu-Raimond et al., 2016; Brovelli et al., 2016; Westrope et al., 2014). The comparison of a region R_1 from a VGI dataset with the same region in an external dataset can yield very different results to the same comparison of a different region R_2 with the same external dataset. Less research attention has been focused on how regions or areas within a VGI dataset or database compare with each other (Davidovic et al., 2016; Barron et al., 2014). The majority of these comparative studies have focused on geometrical comparisons between datasets with a lesser emphasis on comparison of annotations of objects. The research question we are addressing in this paper is as follows: Given two valid regions R_i and R_j from a specific VGI dataset or database how can we measure or assess how similar the two regions are in terms of their annotation of specific object classes? The regions R_i and R_j may be of different geographical size, contain a different number of objects, have different populations, have different numbers of distinct contributors to the specific VGI project, etc. Regions R_i and R_j are separated and neither region is enclosed by

*peter.mooney@maynoothuniversity.ie
†nikola.davidovic@elfak.ni.ac.rs

the other. Practically the regions will represent neighbouring cities, towns, universities, residential areas, etc.

2 Related Work

There is a large body of research results available which describe the comparison of VGI datasets directly with "authoritative" or "gold-standard" datasets or databases (Brovelli et al., 2016; Dorn et al., 2015; Ludwig et al., 2011). However there has been less attention on comparing regions within a specific VGI dataset. In the work of Neis et al. (2013) the authors compare several urban areas in OpenStreetMap. These comparisons are based mainly on contributor activity, density of nodes, ways and relations, data versioning and temporal aspects of contribution. In our recent paper Davidovic et al. (2016) we considered the comparison of annotations of specific objects in OpenStreetMap for 40 cities worldwide and how those cities used the guidance and guidelines from the OpenStreetMap Map Features Wiki. In the work of Barron et al. (2014) a framework containing more than 25 methods and indicators is presented, allowing OSM quality assessments based solely on the datas history to be calculated. Without the usage of a reference data set, approximate statements on OSM data quality are possible. How to compare objects within a VGI dataset is not an easy problem. In their paper Ballatore and Zipf (2015) introduce a concept of *Conceptual Compliance* in VGI which builds a "gold standard" of tags or annotations from the VGI dataset itself and then compares objects against this gold standard. In some cases this can change. Ballatore and Mooney (2015) argue that for VGI contributors to produce meaningful and coherent data these contributors need to negotiate a shared conceptualisation that defines the domain concepts, such as road, building, train station, forest and lake, enabling the communication of geographic knowledge. Henrich and Ldecke (2009) consider the comparing of two different representations of the same geographical region. However the focus in on the geographical footprint of the regions and the degree of similarity or overlap between the two regions in query processing.

3 Performing Similarity Comparisons for Spatial Regions in VGI

We consider the application of the following metrics on the regions R_i and R_j.

- Ballatore and Zipf (2015) introduce their idea of Conceptual Compliance I_{cm}. A set called S which is the source is created containing all of the tags or annotations which are the 'gold standard'. Then for every feature in our region (R) we compute how many of tags or annotations are also in the S source set. This gives us a score between 0 and 1.

$$I_{cm}(R, S) = \frac{\text{count all tags which are both in R and in S}}{\text{Total Number of tags in R}} \tag{1}$$

- Ballatore and Zipf (2015) formalise the concept of Conceptual Richness I_{ri} which is computed as the mean number of attributes or tag keys defined in the features of a region R. This can

be divided into feature classes. The higher the value of I_{ri} the higher the conceptual richness of the VGI dataset is:

$$I_{ri}(R) = \frac{\text{total number of tags used in R}}{\text{The number of features or objects in R}} \tag{2}$$

- Carman et al. (2009) introduce the concept of comparing text vocabularies to see whether the same terms are being used in both. They introduce the idea of *relative overlap* between two vocabularies which might be of different sizes. For a given feature class we extract all of the tags for objects in this class in both regions R_a and R_b. Then the overlap coefficient is computed for two different sets of tags $Tags_a$ in R_a and $Tags_b$ in R_b:

$$Overlap = \frac{Tags_a \cap Tags_b}{min(|Tags_a|, |Tags_b|)} \tag{3}$$

- Using the Pearson Correlation Coefficient. We decided to consider this very well known computation and apply it as follows. If we consider region R_a and R_b as having similarity in annotation. If we consider only the M objects in R_a and R_b which contain some specific tag T (such as tag `highway=primary`). Then we consider the tag keys which occur with our specific tag T in these M objects in both regions. Suppose there are N tag keys $x_0 \dots x_{N-1}$ which are co-occurring with our specific tag T. The Pearson Correlation Coefficient (Puth et al., 2014) is calculated as follows. For the region R_a then for all $x_0 \dots x_{N-1}$ then $R^a_{x_i}$ is the relative usage of the key x_i in R_a, $\overline{R^a_x}$ is the mean of all relative usages of $x_0 \dots x_{N-1}$ in R_a and sR^a_x is the standard deviation of all relative usages of $x_0 \dots x_{N-1}$ in R_a. For example for specific tag T `highway=primary` x_0 might be `name`,x_1 might be `lanes` and x_2 might be `ref`. We can then compute the similarity or correlation or between the two regions R_a and R_b for all annotations of objects with the tag T.

$$r_{R_a,R_b} = \frac{1}{(N-1)} \sum_{i=1}^{N} \Big(\frac{R^a_{x_i} - \overline{R^a_x}}{sR^a_x}\Big)\Big(\frac{R^b_{x_i} - \overline{R^b_x}}{sR^b_x}\Big) \tag{4}$$

4 Results

To provide some initial interpretations of the application of the metrics in Section 3 we extracted all data for the cities of Cambridge, Manchester, Newcastle and Oxford from the OpenStreetMap database on January 2^{nd} 2017. To illustrate the results we chose two very popular target tags $highway = primary$ and $amenity = restaurant$ for our analysis.

Table 1 provides the results of the computation of conceptual compliance for each city for both target tags. The *Objects* column indicates how many objects contained each target tag for each city. We find lower conceptual compliance overall for `highway = primary` objects but overall reasonably high conceptual compliance for `amenity=restaurant`. The high value of conceptual compliance

for Newcastle `amenity=restaurant` indicates that tagging of these objects very closely resembles the gold standard which was taken from the OpenStreetMap editor iD and the OpenStreetMap wiki.

Table 1: Conceptual Compliance calculations for all cities

CITY	**#Objects**	**amenity=restaurant**	**#Objects**	**highway=primary**
Cambridge	146	0.8908	274	0.6202
Manchester	248	0.7165	302	0.7096
Newcastle	166	0.9560	307	0.6694
Oxford	125	0.8842	245	0.6893

In Table 2 the Conceptual Richness is computed for all cities. Ballatore and Zipf (2015) indicate that the higher the Conceptual Richness score the better. The highest value in the table is found for Oxford with `highway=primary`. This indicates a very consistent application of annotations to all objects for this tag.

Table 2: Conceptual Richness calculations for all cities

CITY	**#Objects**	**amenity=restaurant**	**#Objects**	**highway=primary**
Cambridge	146	6.3356	274	6.9330
Manchester	248	4.7218	302	5.2603
Newcastle	166	5.2048	307	4.0199
Oxford	125	6.2661	245	9.7664

In Table 3 and Table 4 the Relative Overlap was computed for `amenity=restaurant` and `highway=primary` for all cities. The highest level of overlap for objects with `highway=primary` is between Newcastle and Manchester (0.7742). The highest level of overlap (0.9032) for objects with `amenity=restaurant` is between Oxford and Newcastle and Manchester and Newcastle. In both cases this is normalised against different numbers of objects in each region. High overlap values indicate that the two regions are annotating objects in similar ways.

Table 3: Relative Overlap - `amenity=restaurant`

CITY	***Cambridge***	***Manchester***	***Newcastle***	***Oxford***
Cambridge	1.0	0.6538	0.8710	0.6042
Manchester	0.6538	1.0	0.9032	0.6042
Newcastle	0.8710	0.9032	1.0	0.9032
Oxford	0.6042	0.6042	0.9032	1.0

Finally the results of computing the Pearson Correlation Coefficient between the cities for `amenity=restaurant` and `highway=primary` are shown in Table 5 and Table 6. For `amenity=restaurant` Oxford and Cambridge are correlated very highly in the tag keys used on objects with this tag. Oxford and Newcastle are also highly correlated. In the case of `highway=primary` the highest correlation values are observed between Manchester and Cambridge and Newcastle and Manchester.

Table 4: Relative Overlap - `highway=primary`

CITY	***Cambridge***	***Manchester***	***Newcastle***	***Oxford***
Cambridge	1.0	0.600	0.6452	0.5660
Manchester	0.6000	1.0	0.7742	0.5742
Newcastle	0.6452	0.7742	1.0	0.6452
Oxford	0.5660	0.5750	0.6452	1.0

Table 5: Pearson Correlation Coefficient - `amenity=restaurant`

CITY	***Cambridge***	***Manchester***	***Newcastle***	***Oxford***
Cambridge	1.0	0.7682	0.8698	0.9402
Manchester	0.7682	1.0	0.8246	0.8014
Newcastle	0.8698	0.8246	1.0	0.9059
Oxford	0.9402	0.8014	0.9059	1.0

5 Conclusions and Future Work

In this paper we have shown some initial results from our efforts to answer the research question on how can we measure or assess the similarity between two regions R_i and R_j from a specific VGI dataset or database in terms of their annotation of specific object classes? The ability to perform similarity comparisons for spatial regions in VGI is important for many reasons. Our focus here has been on comparing annotations of objects within regions in a VGI dataset. Understanding how similar regions are can be helpful for those developing location-based services using VGI, understanding how objects are annotated in those regions, and potentially guiding the VGI project in collection of specific metadata and annotations for object in the region. Four different approaches were proposed in Section 3. There is no broad consensus on which regions are most similar based on the annotation of objects in OpenStreetMap for Cambridge, Manchester, Newcastle and Oxford. Qualitative assessment and evaluation of the results will be required to further investigate the results above. It will be necessary to consider regions R_i and R_j which from anecdotal evidence are known to be either very well mapped or poorly mapped within OSM. It will also be necessary to evaluate the correctness of the values assigned to the keys for the chosen object classes.

There is also the possibility to add comparisons which consider the historical evolution of these regions within a given VGI project (Mooney and Corcoran, 2012). With the focus on comparison of sets of annotations between different regions methodologies outlined by authors such as Rayson and Garside (2000); Basu and Murthy (2015) for discovery of key words in text corpora which can differentiate one corpus from another could be applicable. The biodiversity community also offer some exciting possibilities. Authors such as Boyle et al. (1990); Danilov and Ekelund (1999); Bandeira et al. (2013) outline well known diversity and similarity indices as an approach to estimate biological quality through the structure and abundance of species in communities. We are working to adapt and apply these to tagging and annotation similarity analysis in VGI.

Several authors, such as Neis et al. (2013); Barron et al. (2014) recommend analysis of the number

Table 6: Pearson Correlation Coefficient `-highway=primary`

CITY	***Cambridge***	***Manchester***	***Newcastle***	***Oxford***
Cambridge	1.0	0.8781	0.8474	0.7070
Manchester	0.8781	1.0	0.8584	0.6450
Newcastle	0.8474	0.8584	1.0	0.6289
Oxford	0.7070	0.6450	0.6289	1.0

of contributors in a region, the division of work amongst contributors, etc. as a means of comparing and contrasting the regions. It will be useful to see what influence the number of contributors, contributor work load sharing, etc have when compared with the metrics in Section 3. We shall have additional working results to add to this paper at the time when the reviews of the paper are returned allowing time for these results to be added to the paper for the camera ready version.

Biography

Mr. Nikola Davidovic is a research assistant and a PhD candidate at the Faculty of Electronic Engineering, University of Nis, Serbia. Dr. Peter Mooney is a lecturer and researcher in the Computer Science Department at Maynooth University, Ireland.

References

Ballatore, A. and Mooney, P. (2015). Conceptualising the geographic world: the dimensions of negotiation in crowdsourced cartography. *International Journal of Geographical Information Science*, 29(12):2310–2327.

Ballatore, A. and Zipf, A. (2015). A Conceptual Quality Framework for Volunteered Geographic Information. *Proceedings of COSIT 2015*, pages 89–107.

Bandeira, B., Jamet, J.-L., Jamet, D., and Ginoux, J.-M. (2013). Mathematical convergences of biodiversity indices. *Ecological Indicators*, 29:522 – 528.

Barron, C., Neis, P., and Zipf, A. (2014). A comprehensive framework for intrinsic openstreetmap quality analysis. *Transactions in GIS*, 18(6):877–895.

Basu, T. and Murthy, C. (2015). A similarity assessment technique for effective grouping of documents. *Information Sciences*, 311:149 – 162.

Boyle, T. P., Smillie, G. M., Anderson, J. C., and Beeson, D. R. (1990). A sensitivity analysis of nine diversity and seven similarity indices. *Research Journal of the Water Pollution Control Federation*, 62(6):749–762.

Brovelli, M. A., Minghini, M., Molinari, M., and Mooney, P. (2016). Towards an automated comparison of openstreetmap with authoritative road datasets. *Transactions in GIS*, pages n/a–n/a.

Carman, M. J., Baillie, M., Gwadera, R., and Crestani, F. (2009). A Statistical Comparison of Tag and Query Logs. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 123–130, New York, NY, USA. ACM.

Danilov, R. and Ekelund, N. (1999). The efficiency of seven diversity and one similarity indices based on phytoplankton data for assessing the level of eutrophication in lakes in central sweden. *Science of The Total Environment*, 234(13):15 – 23.

Davidovic, N., Mooney, P., Stoimenov, L., and Minghini, M. (2016). Tagging in volunteered geographic information: An analysis of tagging practices for cities and urban regions in openstreetmap. *ISPRS International Journal of Geo-Information*, 5(12).

Dorn, H., Trnros, T., and Zipf, A. (2015). Quality evaluation of vgi using authoritative dataa comparison with land use data in southern germany. *ISPRS International Journal of Geo-Information*, 4(3):1657–1671.

Henrich, A. and Ldecke, V. (2009). Measuring Similarity of Geographic Regions for Geographic Information Retrieval. In *Advances in Information Retrieval*, pages 781–785. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-00958-7_85.

Ludwig, I., Voss, A., and Krause-Traudes, M. (2011). *A Comparison of the Street Networks of Navteq and OSM in Germany*, pages 65–84. Springer Berlin Heidelberg, Berlin, Heidelberg.

Mooney, P. and Corcoran, P. (2012). The Annotation Process in OpenStreetMap. *Transactions in GIS*, 16(4):561–579.

Neis, P., Zielstra, D., and Zipf, A. (2013). Comparison of volunteered geographic information data contributions and community development for selected world regions. *Future Internet*, 5(2):282–300.

Olteanu-Raimond, A.-M., Hart, G., Foody, G. M., Touya, G., Kellenberger, T., and Demetriou, D. (2016). The scale of vgi in map production: A perspective on european national mapping agencies. *Transactions in GIS*, pages n/a–n/a.

Puth, M.-T., Neuhuser, M., and Ruxton, G. D. (2014). Effective use of pearson's productmoment correlation coefficient. *Animal Behaviour*, 93:183 – 189.

Rayson, P. and Garside, R. (2000). Comparing Corpora Using Frequency Profiling. In *Proceedings of the Workshop on Comparing Corpora - Volume 9*, WCC '00, pages 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Westrope, C., Banick, R., and Levine, M. (2014). Groundtruthing OpenStreetMap Building Damage Assessment. *Procedia Engineering*, 78:29–39.