Stability of Qualitative Spatial Relations between Vernacular Regions Mined from Web Data

Florian A. Twaroch[1], Ross S. Purves[2] and Christopher B. Jones[1]

[1]School of Computer Science, Cardiff University, Wales, UK
{f.a.twaroch,c.b.jones}@cs.cardiff.ac.uk
[2]Department of Geography, University of Zurich, Switzerland
ross.purves@geo.uzh.ch

Vernacular placenames are everyday placenames, which may or may not correspond to administrative gazetteers. We introduce a method to derive a complete set of vernacular placenames for a given region from a listings website, and show how the regions derived can be represented with different geometries. We compared the geometries with regard to computed qualitative spatial relations between the represented regions, and found that the most economical of these geometries, bounding boxes, provide a useful way of presenting information about vernacular regions. 68% of the spatial relations between 27 derived vernacular placenames, determined from bounding boxes, were found to be the same as those obtained with a thresholded kernel density surface representation that incurs significantly higher storage and processing overheads.

1 Introduction

Searches for geographic information on the Internet often fail due to people referring to locations, such as local neigbourhoods, with commonplace or vernacular names that are not recognised by conventional administrative gazetteers [1]. There is a need therefore to acquire knowledge of these place names and record them within gazetteers. The rise of so-called volunteered geographic information [2] on the Internet has introduced a prime source of knowledge of vernacular usage of place names and this has spurred research on mining the Web to obtain knowledge of the properties of vernacular names and to build web-mined gazetteers [3,4,5,6,7]. Volunteered geographic knowledge is found in web pages and associated data that can somehow be linked to geographic coordinates either explicitly, for example in the EXIF data of an image, or implicitly through a place name, address or postcode.

One area of geographic web mining that has received particular attention is the derivation of the borders of vernacular placenames. Initial research concentrated on exploring co-occurrence between a vernacular placename and administrative placenames for which geographic coordinates were available, and then defining the region of the vernacular placename on the basis of these co-occurring point-referenced locations [8]. This work was generally limited to large regions (such as the Alps or the British Midlands) and relied heavily on the quality of geoparsing and

geocoding methods, which are subject to various limitations inherent to methods dealing with unstructured text. Regions were modelled with standard methods such as kernel density estimation and some arbitrary threshold was applied to surfaces thus generated to define a point as being inside or outside of the region.

More recent work has taken advantage of the prevalence of explicitly georeferenced objects now available on the Web (for example, Flickr images and Wikipedia entries) and assumed that placenames used to tag such objects are related to and georefernced by the coordinates assigned [3,5]. Here, the challenges of geoparsing and geocoding are thus largely avoided, and the key problems which remain relate to, firstly, whether tags assigned as placenames do relate to the coordinates, and secondly, methods to assign borders to sets of points representing the same placename.

In general, most of these works have largely focussed on deriving the borders of individual vernacular regions, rather than relationships between them, with the notable exception of Schockaert et al. [7] who mined qualitative spatial relations directly from web documents. In fact, qualitative spatial relations between vernacular regions may provide us with a promising means of storing information about vernacular regions, that can easily be exploited by text-based systems, such as those used in call centres or by car navigation systems. Thus, for example, callers (or drivers) could be provided with the names of a number of regions appropriate to their task, which then act as anchors, containing their location or destination [9]. Importantly, such representations fit well with cognitive models of neighbourhoods, and can be modelled using either standard qualitative spatial relations [10] or fuzzy spatial relations [7].

In this paper we set out to mine, for a given region, the qualitative spatial relations between neighbourhoods used by vendors in an online listings site. We assume, firstly, that neighbourhoods are single continuous regions which have neither holes or disjoint regions associated with them. Secondly, we assume that vendors who assigned coordinates (representing points) and placenames to objects for sale did so in such a way that the coordinates are *contained* by the placename. Finally, we assume that none of the neighbourhoods is *disconnected* from an initial containing region.

Since we are interested in the stability of the qualitative spatial relations derived from the point data, and in methods which are implementable for very large numbers of regions (as would be necessary, for example, in a Geographic Information Retrieval system) we explore how relations vary with different representations of our point set geometry, based on filtered and unfiltered point sets represented by thresholded kernel density surfaces, convex hulls and minimum bounding rectangles.

2 Methods

We mined GumTree[1], a classified ads site, for all georeferenced entries in UK cities. GumTree subdivides its listings according to cities, and we extracted all entries associated with the city of Cardiff. Listings typically have associated with them both a

[1] www.gumtree.com

placename and a coordinate, which we assume to be related. Since this is not always the case, for example where someone from Cardiff advertises a holiday house in Italy, we filter the data for outliers. We do this by, firstly, calculating a surface representing the density of points within a region using kernel density estimation and then retaining only unique points found within the bounding box of the contour enclosing 80% of the volume of the density surface. This simple method appears to effectively filter outliers given our assumption of continuous, singular regions.

For all place names associated with more than five points after filtering, we derived minimum bounding rectangles, convex hulls and the polygon enclosing 80% of the volume of the density surface for the filtered points.

For each of these three geometries we then calculated, for each region with respect to every other region, qualitative spatial relations (QSRs) in terms of the eight possible relations represented by RCC8 (Region Connection Calculus) [10] (Fig. 1).

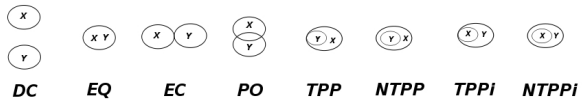

Fig. 1. Qualitative spatial relations in RCC8 (DC: disconnected, EQ: equals, EC: externally connected, PO: partially overlapping, TPP: tangential proper part, NTPP: non-tangential proper part, TPPi: tangential proper part inverse, NTPPi: non-tangential proper part inverse)

3 Results

After removing place names associated with less than five points, a total of 45 unique placenames were identified. Of these 45 unique placenames, 31 were identified through local knowledge as being in Cardiff, and 14 outside. After filtering using the bounding box of the 80% of the volume of the Cardiff density surface, 27 placenames remained, all of which were within Cardiff. Thus, post-filtering, placenames were identified as being in Cardiff with a precision of 100%, and a recall of 87% based on the placenames retrieved from GumTree.

Fig. 2 shows the geometry of the regions identified from the filtered points associated with the 27 placenames for three different geometric representations. These geometries were then used to calculate QSRs.

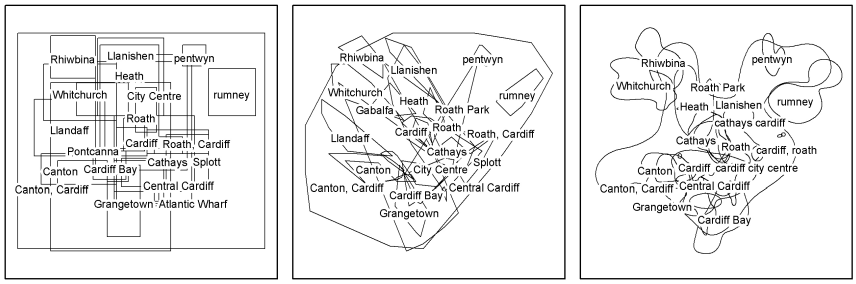

Fig. 2. Geometries for vernacular regions in Cardiff represented using bounding boxes, convex hulls and threshold kernel density surfaces

Table 1 illustrates example QSRs derived for one of the three geometries used with a subset of six placenames. Of the eight possible QSRs in RCC, four are found (PO, DC, NTPP, NTPPi), all of which are illustrated in Table 1. As we also compared regions against themselves to test the robustness of our implementation table 1 contains EQ relationships. Note that we could not find EQ relationships in our data. Also tangential relationships (TPP, TPPi, EC), were not present in any of the derived representations. We retain EQ cases in table one to illustrate that our software would be capable of finding them (EQ, EC, TPP, TPPi) if they ever occur in the input data. A few observations illustrate how QSRs can be used in exploring the relationships between regions. Cardiff, unsurprisingly, contains most other regions, other than Cardiff Bay, with which it has a partial overlap. Roath, Cathays and Canton all appear to be relatively central, having partial overlaps with one another. In contrast, Rumney is isolated from the other regions, though still contained within Cardiff.

Table 1. Example QSRs derived for bounding box geometry for six placenames

Placename	Cardiff Bay	Canton	Cathays	Roath	Cardiff	Rumney
Cardiff Bay	EQ	PO	PO	PO	PO	DC
Canton	PO	EQ	PO	PO	NTPP	DC
Cathays	PO	PO	EQ	PO	NTPP	DC
Roath	PO	PO	PO	EQ	NTPP	DC
Cardiff	PO	NTPPi	NTPPi	NTPPi	EQ	NTTPi
Rumney	DC	DC	DC	DC	NTTP	EQ

To explore the stability of the derived QSRs to changes in the geometric representation of the regions, we calculated contingency tables for QSRs. Table 2 is an example of such a contingency table, showing the change in relations between convex hull and kernel density representations. Values on the diagonal in the table indicate stability - the relation does not change as representations change. Other entries in the table show changes between spatial relationships. Thus, we can see that 58 identical cases of partial overlap were found in both convex hull and kernel density representations, whilst 80 cases which were represented as partial overlaps in convex hulls became disconnected when represented through kernel density. The percentage of QSRs which did not change was 83% for bounding box to convex hull, 79% for

convex hull to kernel density and 68% for bounding box to kernel density respectively.

Table 2. Contingency table changes in QSRs between convex hull and kernel density representation

KDENS / CHULL	PO	DC	NTPP	NTPPi
PO	58	80	4	4
DC	28	456	1	1
NTPP	10	5	17	0
NTPPi	10	5	0	17

4 Discussion

We have developed a method to extract the borders of vernacular regions from GumTree, a dataset where placenames are specifically associated with coordinates. After a filtering step, our method has a precision of 100% within a given region. Recall, within our data is also high (87%), but we do not know how many further placenames are in use in Cardiff. We expect that as volumes of volunteered data increase, so will recall. Techniques which could identify candidate placenames (for example in Flickr tags), would considerably increase the volume of point data available and presumably improve the results of our methods. We treat all placenames returned as vernacular, since we assume here that the use of a placename in GumTree is what defines it as everyday, rather than whether or not it is found in an administrative gazetteer. This assumption is supported by some of our previous work where significant differences were found between the perceived extent of administrative regions and their actual borders [11].

Importantly, our method requires few parameters – a threshold number of points, and density surface volume, as well as standard resolution and bandwidth parameters for the kernel density. The use of convex hulls and bounding boxes removes the need for threshold values to identify regions, and in the case of convex hulls may allow regions to better account for *bona fide* borders to regions, such as railways or rivers bordering a region. Bounding boxes allow for very rapid indexing and computation, making them particularly attractive representations for applications storing very large numbers of regions. Since many production administrative gazetteers are still often point based, they are also an improvement on the state of the art.

Deriving QSRs from the geometries derived is straightforward, and provides a useful way of describing the locations of regions which we believe has considerable potential for a wide range of applications. The relations appear to be relatively stable, though as one would expect there are considerable variations within the QSRs between the different geometries.

5 Conclusions and further work

In this paper we have presented a method for deriving qualitative spatial relations between vernacular regions. Our method requires relatively few parameters, and the initial results, based on mining of a commercial online listing site, are very promising. Qualitative spatial relations provide a powerful way of encoding information about the location of regions in a broader context, using other placenames with which a caller or user of a navigation system may be familiar with as anchors. Furthermore, the relations derived appear to be robust enough with different representations of geometry to suggest that future work should concentrate on collecting larger point sets related to vernacular place names, and on effective methods of outlier filtering.

Acknowledgements
This work was partly funded by the EC FP6-IST 045335 TRIPOD project and by the Ordnance Survey.

References

1. Hill, L. L., Frew, J., Zheng, Q.: Geographic Names. The Implementation of a Gazetteer in a Georeferenced Digital Library, Digital Library. (1999)
2. Goodchild, M. F.: Citizens as Sensors: The World of Volunteered Geography. GeoJournal 69(4), 211--221. (2007)
3. Grothe, C., Schaab, J.: An Evaluation of Kernel Density Estimation and Support Vector Machines for Automated Generation of Footprints for Imprecise Regions from Geotags. In: International Workshop on Computational Models of Place (PLACE'08). Park City, Utah, USA (2008).
4. Jones, C. B., Purves, R. S., Clough, P. D., Joho, H.: Modelling Vague Places with Knowledge from the Web. International Journal of Geographic Information Science 22(10), 1045 -- 1065 (2008)
5. Popescu, A., Grefenstette, G., Moëllic, P.: Gazetiki: Automatic Creation of a Geographical Gazetteer. JCDL 2008: pp. 85--93. (2008)
6. Schockaert, S., Cock, M. D.: Neighborhood Restrictions in Geographic IR. In: SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA, ACM Press, pp. 167--174. (2007)
7. Schockaert, S., Smart, P. D., Abdelmoty, A., Jones, C. B.: Mining Topological Relations from the Web. In: DEXA Workshops 2008, pp. 652--656. (2008)
8. Purves, R., Clough, P., Joho, H.: Identifying Imprecise Regions for Geographic Information Retrieval Using the Web. In: GIS Research UK 13th Annual Conference - GISRUK 2005, Glasgow, pp. 313--318. (2005)
9. Hood, J., Galton, A.: Implementing Anchoring. In: M. Raubal, H. Miller, A. Frank, and M. Goodchild (editors), Geographic Information Science: Fourth International Conference, GIScience 2006, Münster, Germany, September 2006, Springer, pp. 168 --185. (2006)
10. Randell, D. A., Cui Z., Cohn, A.G.: A Spatial Logic based on Regions and Connection. KR 1992: 165--176 (1992)
11. Twaroch F.A., Jones, C.B. and Abdelmoty, A.I.: Acquisition of a Vernacular Gazetteer from Web Sources. In: Workshop on Location on the Web 2008, Locweb 2008, Beijing, China, pp. 61--64. (2008)