Acquisition of a Vernacular Gazetteer from Web Sources

Florian A. Twaroch
Cardiff University,
School of Computer Science
Cardiff, UK
+44 2920 876058
f.a.twaroch@cf.cs.ac.uk

Christopher B. Jones
Cardiff University,
School of Computer Science
Cardiff, UK
+44 2920 874796
c.b.jones@cs.cf.ac.uk

Alia I. Abdelmoty
Cardiff University
School of Computer Science
Cardiff, UK
+44 2920 874751
a.i.abdelmoty@cs.cf.ac.uk

ABSTRACT

Vernacular place names are names that are commonly in use to refer to geographical places. For purposes of effective information retrieval, the spatial extent associated with these names should be able to reflect people's perception of the place, even though this may differ sometimes from the administrative definition of the same place name. Due to their informal nature, vernacular place names are hard to capture, but methods to acquire and define vernacular place names are of great benefit to search engines and all kind of information services that deal with geographic data. This paper discusses the acquisition of vernacular use of place names from web sources and their representation as surface models derived by kernel density estimators.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms

Algorithms, Measurement, Experimentation, Human Factors

Keywords

vernacular geography, web mining, geographic information retrieval

1. INTRODUCTION

Place names play a key role in formulating queries for geographical information retrieval on the web. A typical generic structure for explicit enquiries about geographical information takes the form of triplets of <*subject*><*relation*><*somewhere*> in which the subject specifies the thematic aspect of the search, the somewhere is the name of a place and the relation stipulates a spatial relationship to the named place such as "in", "near" or "north of". Processing a query in this form usually entails transforming the place name and its qualifying spatial relation to a query footprint that represents a region of space to which the query is assumed to refer. Generation of a query footprint requires that the place name itself is represented by a footprint which is then modified according to the spatial relation. For the purposes of most geographical web search facilities, gazetteers provide the main source of knowledge of the footprint associated with place names. Place footprints are frequently just a single point, but they may also be a bounding box or a polygon. The majority of gazetteers are derived from the content of topographic maps produced by national mapping agencies and as such they represent a relatively "official" or administrative view of geography. This causes a problem for the geo-information services that use these gazetteers because people often use vernacular place names that are not recorded in the gazetteers and hence result in failure to process a query that contains such a name.

Vernacular place names are names that are commonly in use to refer to geographical places and the spatial extent associated with them reflects the common perception. In many cases, as indicated above, the name may not correspond to an officially designated region or place. Examples would be the South of France, the English Midlands and the American Mid West. Many vernacular names, such as these, are vague in spatial extent. Thus there may be locations (possibly corresponding to other named places) that most people would agree are part of the vernacular place and others that are borderline without uniform agreement. Sometimes a vernacular name may be the same as an official name but the common understanding of its spatial extent may not match exactly with the official interpretation. For public access information systems the objective is to understand what the user is referring to and so it is the vernacular interpretation of a place that is required in order to meet the user's needs.

We are faced then with a challenge to acquire knowledge of the intended spatial interpretation of vernacular place names. There have been several earlier descriptions (summarised in the next section) of techniques for acquiring vernacular place name knowledge that are relatively labour intensive. More recently it has become apparent that the web itself is a valuable source of such knowledge. It has been observed that web pages that include a vernacular place name often include the names of other places that are inside or in the vicinity of the vernacular place. Maps of the extent of the vernacular places can be generated from the locations of the most frequent co-occurring names. Here we exploit just a few individual web resources that contain numerous geo-referenced place names that relate to business entities and to other private or community services and facilities. One of these sources, Google Maps, enables retrieval of the coordinates of businesses, georeferenced by their address, and other "community" entities with vernacular place names that have been geo-referenced in Google Earth. Another source, the Gumtree website, enables retrieval (screen scraping) of the georeferences of advertised services for which a place name has been provided.

LocWeb 2008, April 22, 2008, Beijing, China.

In what follows we review briefly previous efforts to acquire knowledge of vernacular places before describing how georeferences of vernacular names are extracted from the selected web sites. We then present methods for visualising and modelling the spatial extent of the extracted point clouds that are associated with individual place names. We discuss the relative merits of the different sources that we employ and analyse sources with different bias. The paper concludes with proposals for future work.

2. STATE OF THE ART

The present work proposes models of vernacular place name geography on a nationwide scale for which automated methods to acquire the relevant data are required. Recent efforts to model vernacular place names such as 'downtown' are based on human subject tests and interviews (cf. [4]) but turn out to be too labour intensive for the definition of vernacular place names for a whole country.

Automated definition of city centres in the UK has been based on census and socioeconomic data. The latter served to derive indices for property, economy, diversity and visitor attractions. Each index has been modelled as a density surface model and combined with map overlay operations to yield a surface model of 'town centeredness' [11]. A comparison of how the derived representation matches people's cognition of city centres was not provided.

A web based method that considers the cognition of place has been implemented by [2]. The authors utilized a spray can tool to define high crime regions in Leeds. A spray can allows users to define vague regions through drawing clouds of different point density on a map and label the sprayed contents accordingly. The tool also allows one to define crisp boundary features by spraying hard edges, and to distinguish between locations that are better examples for a certain place than others by spraying more in certain locations of a map than at others. This is in accordance with the typicality concept in cognitive science [7], stating that some locations might be better examples for a certain region than others. However the tool is biased by the maps used and needs to be tested at different scales.

Vernacular place names are often vague in their nature and crisp definitions do not exist. Fuzzy footprints have been defined utilizing trigger phrases and other web queries to search for places that lie within a region under consideration, regions that include the region under investigation and regions that are neighbouring the investigated region [8]. The derived place names have been grounded with the Alexandria Digital Library gazetteer. The study was carried out on political regions in order to validate the achieved results. Schockaert and Cock state in a later paper [9] that the perception of administrative boundaries often deviates from the 'official' definition.

Pasley et al [6] investigate the definition of imprecise regions of different sizes using web queries and a geo-tagging algorithm. The study reveals that regions as big as several counties compared to vernacular place names in city environments have to be treated differently as the source of error in geo-tagging changes with the scale and the used resources.

3. EXTRACTING VERNACULAR KNOWLEDGE FROM WEB SOURCES

One measure of usability of current web systems is in the extent to which users can express queries using place names that reflect vernacular geography and then gain access to the relevant resources effectively and efficiently. Progress towards this goal may be achieved by complementing the traditional gazetteer services with gazetteers of vernacular place names, populated from web resources.

Multiple sources of vernacular place names are emerging on the web. In this paper we focus on social web applications as a potentially rich source for collating this information. Web sites such as Flickr[1] and Geograph[2] are facilitating the geo-tagging of personal resources, allowing people to markup photo collections. Place vocabularies used on those sites include place names and relationships to place names. For this study, we focus mainly on studying absolute references to places and their location. Another category of web systems offering structured place information are yellow pages and other business directories. Specific examples from both types of resources are used here to illustrate the study.

For illustration purposes, the paper focuses on a specific geographic region of Cardiff, Wales, UK. The authors are familiar with the study area and can identify gross errors. We briefly describe sources that allow the constraining of search to a specific geographic region.

3.1 Geo-References from Social Websites

The social website Gumtree[3] serves a community to trade items and properties as well as offer a virtual place to meet and arrange meetings in real space. A free ad can be posted on the website. Users can associate the ad with a postcode which can then be published by Gumtree using a Google map service to display the location of the ad. Place name data on this site have been mined to find vernacular regions in the city of Cardiff with the aim of finding clusters of points labelled with place names. Figure 1 shows an example of a map of a point cluster located in a region of Cardiff called "Roath".

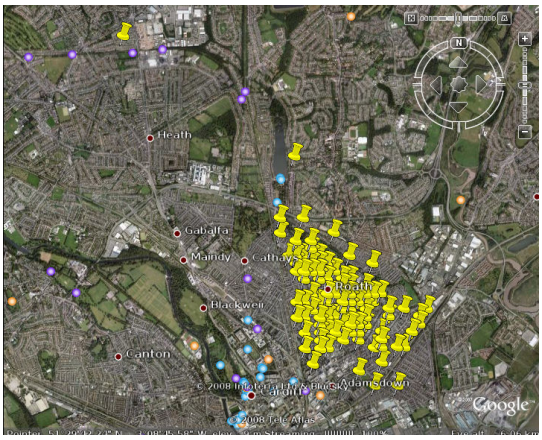

Figure 1: Points mined from Gumtree labelled "Roath"

[1] http://www.flickr.com

[2] http://www.geograph.org.uk

[3] http://www.gumtree.com

The density of the mined data points is dependent on the availability of "current" ads related to specific regions. In some regions only a few labelled points could be mined from Gumtree. Hence, mining other resources in a similar way is needed to increase the density of the data collected. Recently, Google offers the possibility to create maps by manually placing markers on a map and sharing them on the web. Many other georeferenced data (e.g. kml files of gps tracks of hiking and cycling tours) have been additionally indexed and can be found through the Google maps search engine known as 'community maps'. An example of mining this data is shown in the right image in figure 2.

3.2 Geo-References from Business Directories

Google Maps offers a free service called Local Business Center where businesses can register their location with some descriptive contents and are in turn indexed by Google's search engine, showing up on Google's map service. In the spirit of [12] we assumed that place names can occur as part of business names, such as those registered with Google. We sent queries of place names found in Gumtree to Google's map search engine and mined the results retrieved through these queries (Figure 2 left).

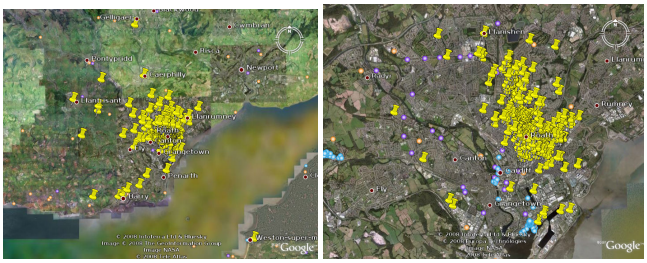

Figure 2: Points mined from Google (left – Business Directory, right – Community Maps) associated with "Roath"

Visual inspection of the point clouds reveals that the point data mined through Google's service (figure 2) are much more scattered than those mined from Gumtree (figure 1). The present mining algorithm does not consider spatial relationships. While data from Gumtree is always meant to lie *in* the investigated region, data mined through Google may be also located *near, around, etc.* the region of interest. The scattered data mined through Google business maps also suggests that businesses can carry place name information far away from the actual place.

4. GEOGRAPHICAL REPRESENTATION OF VERNACULAR PLACE NAMES

The described mining methods facilitate retrieval of a huge amount of point data for place names so that we can apply methods from spatial statistics, specifically kernel density estimation [10] to represent the place names. In the following two subsections we briefly describe how outliers that can skew the data have been identified and introduce the kernel density estimation method.

4.1 Outlier Identification

We constrained the mined point data to a certain geographic region and can therefore eliminate coordinates not being located within the boundary polygon of the superior region of the place names under investigation, i.e. the city of Cardiff.

Multiple postings of a place name by a single person can falsify the result by skewing the data to a single point. We apply a simple heuristic to get rid of multiple postings: Both, markers placed by hand or positioned by GPS exceed a certain measurement error. We can therefore delete points within an epsilon region of the measurement error. So we can remove duplicate data mined from Google before applying the kernel density estimation.

4.2 Kernel Density Estimation

Kernel density estimation has been applied in the literature [3, 11] to represent vernacular place name geography. The principle of KDE is based on determining a weighted average of data points within a moving window centred on a grid of points p. KDE turns a vector into a field representation. Different kernel functions can be applied, but it has been previously found [5] that the choice of the kernel function is less important than the choice of the bandwidth parameter. This parameter controls the influence of the kernel functions on the summed local intensity values. As we investigate regions within a city environment we set the parameter to 300 meter (cf. [11]). We are aware that at this point we will have to improve our method by investigating adaptive methods such as those proposed by [1] that allows defining and adapting the bandwidth of the kernel based on the underlying data automatically. The kernel function used in this paper is a "Gaussian kernel" and the results presented have been produced with GRASS GIS[4].

4.3 Current Results

The present method is suitable for modelling the extent of place names in populated places, especially city environments. Based on the evaluation of the retrieved data sets we can classify three types of place names: 1) Place names whose commonly perceived extent coincides with the administrative definition. 2) Place names whose extent does not coincide with an existing definition. 3) Place names that exist in people's minds but not in the administrative geography.

The third type can be validated by comparing representations achieved through other independent methods such as questionnaires. As this implies a number of factors we have not considered yet, we concentrated first on places that have been defined by administrative authorities.

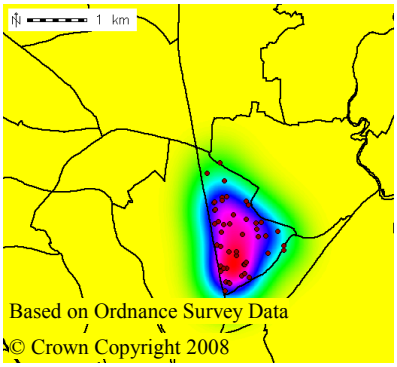

Figure 3. Vernacular and administrative definition coincide (Plasnewydd – ward in Cardiff)

Figure 3 shows an example where the derived region coincides with the administrative geography. All points derived from Google community maps are within the boundary of the

[4] http://grass.itc.it

administrative definition, suggesting that people's use of the place name coincides with its administrative definition. The name does not seem very popular as neither Gumtree nor Google business queries yield enough data points to derive further representations.

Mining data for the neighbouring region "Roath" reveals that people's cognition of "Roath" differs significantly from the administrative geography (see figure 4). Data from Gumtree even suggests that the former place Plasnewydd is overridden by the definition of "Roath" in people's mind. A possible explanation for this result is that the region "Roath" is a popular area where students and families with children are living. A number of web documents that promote real estate would therefore refer more often to "Roath" than to less popular adjacent areas. Future research will uncover such effects by mining and analysing further data from the web sources such as the author's identity, the intention of the description, the age of the data source and others.

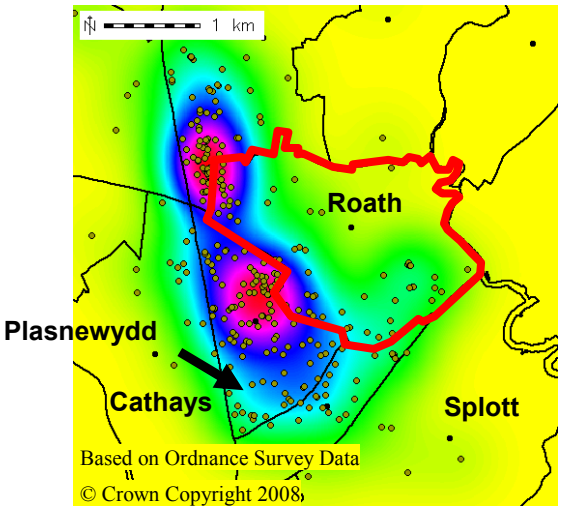

Figure 4. Vernacular and administrative definition do not coincide (Roath– ward in Cardiff)

We found that the phenomenon of points being relatively highly scattered, as observed above for the area of Roath, applied in general for data derived from business maps. The investigated regions represented by business data differed considerably in size compared to regions derived solely from community driven data. Here further research is necessary and other data sources such as Yellow page data and place name data derived through web questionnaires should be considered. Sound methods to combine data from different sources will then be required.

Place names that exist in people's minds but not in administrative geography have to be investigated further. "Cardiff Bay" is an example for such a place name and first experiments (not presented here) show promising results. Methods to validate the derived results are necessary.

5. Conclusions and Future Work

The proposed method is capable of representing the spatial extent of vernacular place names by querying the web, and we have shown that the resulting representations can differ significantly from the equivalently named administrative geography. A priority for future work is the validation of the results. Future work will also have to deal with combining data from different (web) sources. We will therefore carry out a web questionnaire on a large scale to gain access to an independent data source.

The identification of vernacular place names within a query is a problem in itself that has not been addressed in the present paper. It requires to identify a term as being a place and a method to measure the degree of vernacularity

For the integration of different web sources there is a need to define quality measures for the derived representations. Further methods to mine data from the web and derive representations according to the vernacular geography of the UK will be investigated.

6. ACKNOWLEDGMENTS

We would like to thank Ordnance Survey for funding our research on representation of place for geographic information retrieval. This work has also been partly funded by the EC FP6-IST 045335 TRIPOD project.

7. REFERENCES

[1] Brunsdon, C., Estimating probability surfaces for geographical point data: an adaptive kernel algorithm, in Computers & Geosciences. 1995, Pergamon Press, Inc. p. 877-894.

[2] Evans, A.J. and T. Waters, Mapping Vernacular Geography: Web-based GIS Tools for Capturing 'Fuzzy' or 'Vague' Entities, in International Journal of Technology, Policy and Management. 2007. p. 134--150.

[3] Jones, C.B., et al., Modelling Vague Places with Knowledge from the Web, in International Journal of Geographic Information Systems.

[4] Montello, D.R., et al., Where's downtown?: Behavioral methods for determining referents of vague spatial queries, in Spatial Cognition and Computation. 2003. p. 185--204.

[5] O' Sullivan David and Unwin, D.J., Geographic Information Analysis. 2002: Wiley.

[6] Pasley, R., P. Clough, and M. Sanderson, Geo-Tagging for Imprecise Regions of Different Sizes, in Proceedings of Workshop on Geographic Information Retrieval GIR'07. 2007.

[7] Rosch, E., Cognition and Categorization, E.L.B.B. Rosch, Editor. 1978, Lawrence Erlbaum Publishers. p. 27-48.

[8] Schockaert, S., M.D. Cock, and E.E. Kerre, Automatic Acquisition of Fuzzy Footprints, in OTM 2005 Workshops (SeBGIS 2005). 2005. p. 1077-1086.

[9] Schockaert, S. and M.D. Cock, Neighborhood restrictions in geographic IR, in SIGIR '07. 2007, ACM Press: New York, NY, USA. p. 167-174.

[10] Silverman, B.W., Density estimation: for statistics and data analysis, ed. Chapman and Hall. 1986, London.

[11] Thurstain-Goodwin, M. and D. Unwin, Defining and delineating the central areas of towns for statistical monitoring using continuous surface representations, in Transactions in GIS. 2000. p. 305-317.

[12] Zelinsky, W., North America's vernacular regions, in Annals of the Association of American Geographers. 1980. p. 1-16.