Acquisition of Vernacular Place Names from Web Sources

Florian A. Twaroch, Christopher B. Jones and Alia I. Abdelmoty

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 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 kinds 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. We show that various web sources containing user created geographic information and business data can be used to represent neighbourhoods in Cardiff, UK. The resulting representations can differ in their spatial extent from administrative definitions. The chapter closes with an outlook on future research questions.

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.

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 triples 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 modi-

fied 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 (lower left and upper right coordinate of a rectangle) 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 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 Midwest". 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 a user is referring to and so it is the vernacular interpretation of a place that is required in order to meet users' needs.

Consequently, we are faced with the 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 (Purves et al. 2005). Maps of the extent of vernacular places can be generated from 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 user created entities with vernacular place names. Another source, the Gumtree web site, has been screen scraped to acquire georeferences of advertised services for which a place name has been provided.

In what follows we review briefly previous efforts to acquire knowledge of vernacular places before describing how georeferences of vernacular names are extracted from 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 chapter concludes with proposals for future work.

2. REPRESENTATION OF PLACE

The perception and cognition of place varies among people. Even for the same person a place might be differently perceived given two different contexts. In this section we review literature to represent place formally. Formal representations of place may help to improve the interfaces of information and decision support systems. Our aim is to use automated methods to build representations of vernacular place name geography on a nationwide scale.

A method to represent vernacular place names such as "downtown" has been based on human subject tests and interviews (cf. Montello et al. 2003), but turns 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" (Thurstain-Goodwin and Unwin 2000). 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 Evans and Waters (2007). 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 (Rosch 1978), stating that some locations might be better examples for a certain region than others. However, the regions captured with such a tool are biased by the maps used and Evans and Waters (2007) did not report tests on a wide scale.

The neighbourhood project is another example for a web based tool to capture vernacular geography (http://hood.theory.org). It is a mashup of Craig's List and Google Maps. People are asked to provide a postcode and a place name. The postcodes are converted into coordinates utilising a postcode database. The resulting point clouds are analysed using a metaball algorithm to define clusters, i.e. neighbourhoods (Geiss 2000).. As this method is not statistically grounded, two problems appear: 1) multiple dense points cause the algorithm to overemphasize

the spatial extent of neighbourhoods and 2) the method can not deal well with outliers.

Recently a considerable number of researchers have used search engines to query the Web as a source from which to extract information to model vernacular place names (Purves et al. 2005; Schockaert and Cock 2007; Jones et al. 2008). In these approaches, web pages are parsed automatically for references to places. Some of these places can be found in gazetteer services like the Alexandria Digital Library (Hill et al. 1999) or the Getty Thesaurus of Geographic Names (Harpring 1997). Using the gazetteer services, these places can be grounded, i.e. georeferenced with a set of coordinates. Borges et al. (2007) carried out experiments to evaluate the presence and incidence of urban addresses in web pages to discover geographic locations. A set of web pages is collected using e.g. a web crawler. In a *geoparsing* step, potential geographic entities are identified such as postcodes, telephone numbers, and other address information, and converted into structured addresses to feed a *geocoding* process, i.e. looking the address up in a gazetteer. The authors tested 4 million pages and found that, in 15% of the web pages, addresses could be found. Furthermore, they concluded that postal codes were superior over other address information as a geocoding resource.

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 (Schockaert et al. 2005). 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 (Schockaert and Cock 2007) that the perception of administrative boundaries often deviates from the "official" definition.

Other work by Schockaert et al. (2008) addresses the detection of place names for the region of Cardiff, UK. The authors use a focused crawler to extract relevant web pages, similar to Borges et al. (2007), and interpret addresses found on web pages. Heuristics based on rules and document frequencies serve to define filters. With the filtered list of place names spatial relationships are extracted, and their consistency is tested with further queries to the Web and fuzzy spatial reasoning.

Pasley, Clough, and Sanderson (2007) 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 have to be treated differently from vernacular place names in city environments, as the source of error in geotagging changes with the scale and the used resources.

Whether or not vague phenomena can be described at all by a crisp boundary polygon is currently an open question (Evans and Waters 2007). A number of qualitative representation methods have been proposed in recent literature (Bennett 2001; Kulik 2001; Vögele et al. 2003).

A simple way to describe a region utilizing a set of points, labelled as belonging to a specific region has been summarized by (Galton and Duckham 2006).

Alani et al. (2001) and Arampatzis et al. (2006) describe methods based on Voronoi diagrams and Delaunay triangulations to determine approximate boundaries of regions represented by sets of points.

A representation of a vernacular place should maintain the uncertainty of the definition. Different people have different beliefs about how to define the extent of a certain place. Methods based on fuzzy regions (Schockaert et al. 2005; Schockaert and Cock 2007) consider that fact in providing models of spatial regions carrying a varying degree of membership.

Our work is similar to previous methods found in the literature (Purves et al. 2005; Jones et al. 2008). The delineation of the spatial extent of vernacular place names is based on kernel density estimation methods. Thresholds of these models at different levels yield footprints of certain confidence values, expressing degrees of familiarity with the modelled region. See also Chapter 9 of this volume on the determination of geographic representations for arbitrary concepts at query time.

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 and Geograph are facilitating the geo-tagging of personal resources, allowing people to annotate photo collections. Other sites such as Gumtree allow local sourcing of products and services in an informal way; Wikipedia is a user maintained web-based encyclopaedia with extensive geographic coverage (Overell and Rüger 2007); Placeopedia is a project to geocode Wikipedia articles. In summary, a plethora of volunteered geographic information exists on the Web and the amount of available data is growing every day (Goodchild 2007).

Place vocabularies used on volunteered geographic sites include place names and relationships to place names. For this study, we focus mainly on studying absolute references to places and their location in web resources, in that the vernacular names are accompanied by an explicit geographic reference such as a postcode. This is in contrast to earlier work, such as that of Purves et al. (2005) or Jones et al. (2008), in which the extent of vernacular regions is found indirectly in terms of the georeferences of places that are assumed to lie inside the vernacular place. The current approach makes the process of acquiring vernacular names data much

simpler and eliminates errors introduced during the extraction (and grounding) of geo-references.

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.

We have queried these latter sites using a web mining method (explained in Section 3.1), applying two different strategies to mine data for the representation of vernacular regions. Firstly, we utilize businesses addresses to delineate the spatial extent of vernacular regions (Section 3.2). Then we query sources of user created geographic information, such as social web pages and community directories, to gain access to user generated content (Section 3.3). For illustration purposes, the paper focuses on regions within the broader geographic region of "Cardiff, UK". The authors are familiar with the study area and can identify gross errors in the retrieved vernacular data.

3.1 Mining the Web

Search engines are a good source of data since their aim is to index any web page that might be of interest to some audience. Search engines rely on software agents to discover pages automatically by following links in known pages. They do so by processing an initial seed list of Uniform Resource Identifiers (URIs) and crawling the URIs recursively from there. Pages discovered are stored in an index which can be searched using a user interface. The search engines generally also expose search services in developer friendly ways that allow the data to be collected (e.g. through an Application Programming Interface, or API).

A large part of the Web is not indexed and search engines themselves introduce biases of unknown nature. Secret ranking algorithms rate and interpret the content of web pages. The Web which exists but is not indexed by search engines, can be termed the "Hidden Web" (Ipeirotis et al. 2006). Often it contains databases that can be accessed via web pages. A technique to collect data from this set of pages is called web scraping. Web scraping can be utilized whenever no public API is available. A web scraping program issues automatically created URIs to query the database behind web pages. The program can automate tasks such as filling in forms. A crawler cannot find such pages as it always seeks for links but does not interact with a web page. For example to query an imaginary web page the following URI might be used http://www.example.com?q=cardiff&page=3. In this hypothetical case the third page about Cardiff would be returned. Since this would be returned in a format designed to be rendered by a browser, further processing would then be needed to extract information as required. This would normally use pattern matching or Natural Language Processing (NLP). It can be seen that this requires knowledge about the structure of the URI and the markup format of the

web pages. Whenever the provider of the site changes structure or contents, the scraping tool needs to be adapted.

Data sets mined with web scraping tools can be considerably different from data sets retrieved via the exposed query APIs. APIs are often restricted in their content and do not allow to extract all the relevant information (Frank McCown and Nelson 2007). Therefore a web scraping tool has been implemented that creates URIs to query Google Maps. The required parameters are documented on Wiki pages for the Google Maps API (Mapki 2008). The most important parameters are listed below:

- q: a query string
- near: a place name that the returned point will be related to
- start: an integer. The results returned will be those found after ignoring this many results.
- num: return at most this number of matches
- mrt: a parameter that speficies the search type
    - mrt=loc: specifies a location search.
    - mrt=yp: searches for businesses, yp stands for yellow pages.
    - mrt=kmlkmz: data indexed as "user contributed contents"

Queries were submitted with URIs containing "city centre" or "Roath" (a community in Cardiff) in the q term and "Cardiff" in the near term. Spaces and special characters have to be escaped in the URIs. The number of matches is by default set to ten (num=10). In order to receive the next 10 results, a new query had to be formulated with the start value incremented by 10. We generated queries until the start value reached 200. One series of queries issued collected just business data (e.g. http://maps.google.co.uk/?q=city%20centre&near=cardiff&mrt=yp) another series just user contributed contents (e.g. http://maps.google.co.uk/?q=city%20centre&near=cardiff&mrt=kmlkmz).
Using the various parameters described, a set of web pages was collected. Placenames and coordinates were retrieved from the pages using natural language processing techniques.

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 Zelinsky (1980) 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 (see Section 3.1) and mined the results retrieved through these queries.

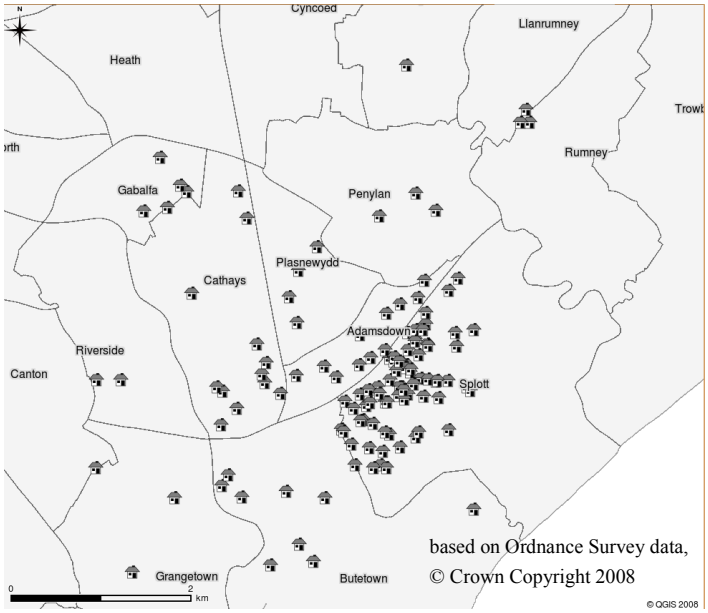

Fig. 1. Points mined from Google's Business Directories associated with "Splott" (a ward and community in "Cardiff"),

The postcode of each business can be looked up in a postcode database and facilitates geocoding. Each business can then be associated with coordinates and visualized on a map (Fig. 1). This step was not necessary as the coordinates for each business could be directly found in the mined web pages.

Fig. 1 illustrates the results mined for the administrative area "Splott", a ward and community in "Cardiff". The mined data points, symbolized by small houses, are scattered over more areas than just the one labelled as "Splott". The reason for that is partly due to the methodology used.

Businesses such as real estate agents that are actually not located in the ward "Splott", but refer to it, are currently not filtered out. The scattered data mined through Google business maps suggests that businesses can carry place names, as part of the business name, relating to places that are far away from the location of the business. Future versions of the present mining algorithm will have to consider spatial relationships mentioned on business web pages and consider if the mined places are located "near", "around", etc. the region of interest.

3.3 Geo-References from a Social Website

The social web site Gumtree serves a user-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 web site. 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. Fig. 2 shows an example of a map of a point cluster located in the ward "Splott" in "Cardiff".

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 over a longer period of time is needed to increase the density of the data collected.

The different coloured triangles in Fig. 2 represent data from two different periods of time associated with the label "Splott". Each triangle represents an ad in Gumtree. Fig. 2 shows that the majority of mined data points lie within or near "Splott," suggesting that data mined through the social web site are more suitable for direct geocoding than data coded through business addresses (Section 3.2).

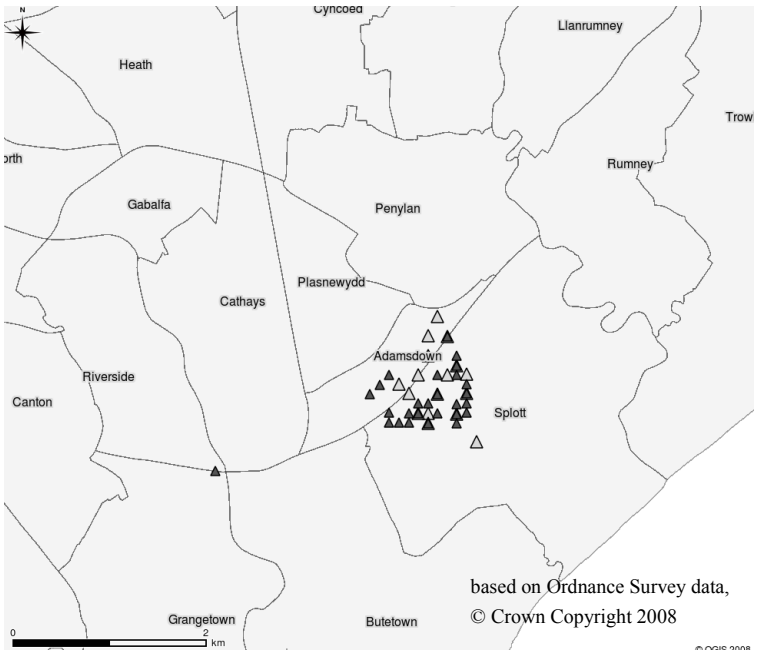

Fig. 2. Points mined from Google's user created contents vs. points mined from Gumtree associated with "Splott"

3.4 Geo-References from Google User Created Geographic Information Sources

In order to increase the point density for regions, other sources of volunteered geographic information (Goodchild 2007) have to be queried. This would require writing a number of web scraping tools for each of the sources. Instead, we utilized a single web scraping tool via Google's search engine. Recently, Google indexed a number of user created geographic information sources e.g. KML files of GPS tracks of hiking and cycling tours, geotagged photos, etc. and made them available through the web interface of their search engine (http://maps.google.co.uk/maps – choose user created contents as search option, formerly Google community maps). Since then the company also offers the possibility to create maps by manually placing markers on a map and sharing them on the web, and tools to provide feedback about the correctness of a marker placed on a map. KML is an OGC Standard based on an XML schema to geographically annotate and visualize data on the Web. KML files carry geometric information such as point, line or polygon data but can have a variety of other additional attributes, especially place name labels. These place name labels have been utilized to associate place names with points and model the regions of Cardiff. A number of items found through this interface actually referred to business related items again.

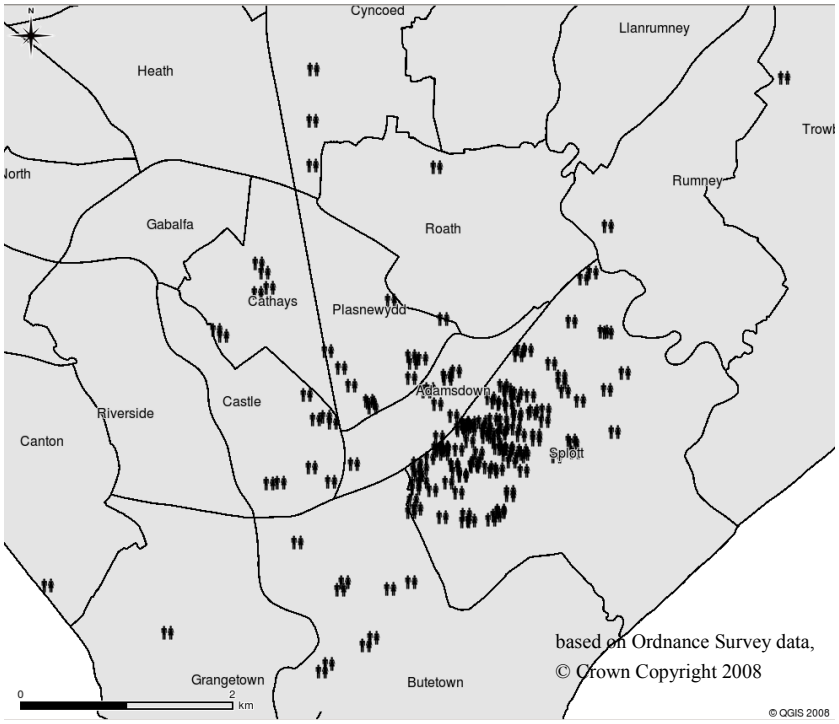

Fig. 3. Data points for the ward "Splott" mined from different sources that were indexed by Google as user created contents.

Fig 3. shows that the majority of points mined from data sources that Google describes as user created content coincide with the region of interest. Mining more of these sources, i.e. Google user created contents, introduces more noise. A number of points, similar to the result for data points mined from business directories, are found to lie outside and far away from the actual region. Future work will have to address the development of quality measures to judge the accuracy of different Google user created web sources.

4. DETERMINATION OF THE SPATIAL EXTENT OF VERNACULAR PLACE NAMES

The described mining methods facilitate retrieval of a huge amount of point data for place names. We can then apply methods from spatial statistics, specifically kernel density estimation (Silverman 1986), to represent the spatial extent of 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 that are not 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 simple heuristics to get rid of multiple postings: Markers placed by hand or positioned by GPS differ from the multiple posting data in that they exceed a certain measurement error. We can therefore delete points within an epsilon region of the measurement error and hence 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 (Thurstain-Goodwin and Unwin 2000; Jones et al. 2008) 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 (O' Sullivan and Unwin 2002) 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 300m (cf. Thurstain-Goodwin and Unwin 2000). We are aware that at this point we will have to improve our method by investigating adaptive methods such as those proposed by Brunsdon (1995), that allow automatic defining and adapting of the bandwidth of the kernel based on the underlying data. The kernel function used in this chapter is a “Gaussian kernel”.

4.3 Current Results

In our approach we use coordinates derived from geotags. Three different sources are available and thus each region of Cardiff can be modelled three times using the data points from the three different sources. We compare the different

results by visual inspection and evaluate them using local expert knowledge. Based on this evaluation, three types of place names could be classified: 1) place names whose commonly perceived extent coincides with the administrative definition of the same name, 2) place names whose extent does not coincide with the correspondingly named administrative definition and 3) place names that exist in people's minds but not in the administrative geography.

4.3.1 Administrative Places Names

Part of our experimental work was to find out if people's description of certain places coincides with administrative definitions. For a number of wards we found that data points mined and models derived approximate the spatial extent of the investigated region as defined by the available administrative boundary datasets.

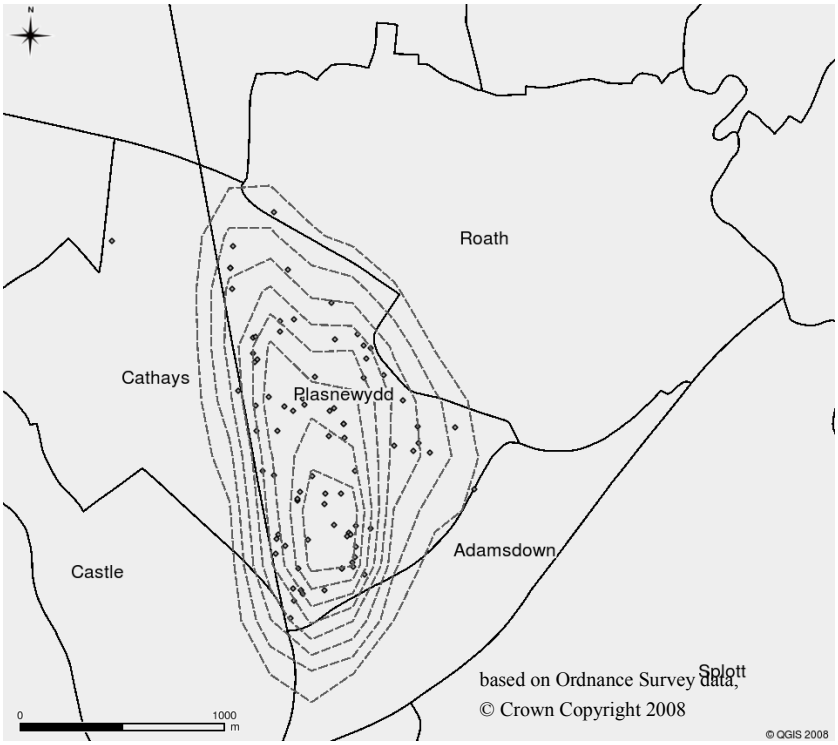

Fig. 4. Vernacular and administrative definition coincide ("Plasnewydd" – ward in "Cardiff")

Fig. 4 shows an example where the derived region coincides with the administrative geography. Almost all points derived from Google user created contents are

within the boundary of the administrative definition, suggesting that people's use of the place name coincides with its administrative definition. The name ("Plasnewydd") does not seem very popular as neither of the other data sources, i.e. Gumtree and Google business queries, yield enough data points to derive further representations.

4.3.2 Semi-Vernacular Place Names

People's perception of the spatial extent of a place can significantly deviate from the administrative definition. When mining data from the three different sources for the region "Roath", which neighbours "Plasnewydd" in the administrative geography, we found that the majority of points were actually not located in the community labelled as "Roath" (Fig. 5). Data from Gumtree even suggests that the former place "Plasnewydd" is overridden by the definition of "Roath" in people's mind. Data points from Gumtree that should be labelled as "Plasnewydd" were consistently labelled as "Roath".

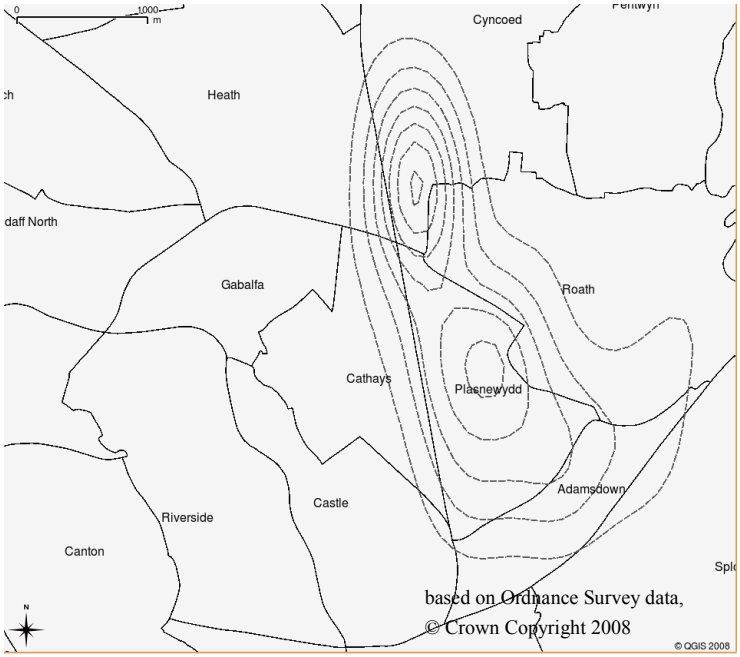

Fig. 5. Vernacular and administrative definition do not coincide

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 docu-

ments 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.

4.3.3 Vernacular Place Names

Some place names like "city centre" do not exist in administrative geography. A new development area in Cardiff is called "Cardiff Bay," and its spatial extent has not been defined in administrative boundary data sets. We used the three different data sources to create a model of "Cardiff Bay" (Fig. 6).

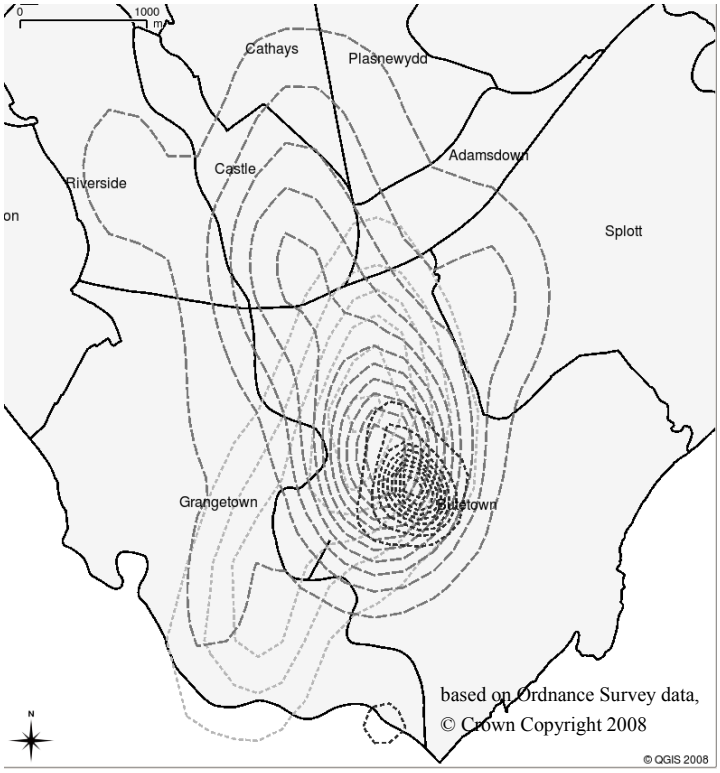

Fig. 6. Three different kernel density estimates for an area known as "Cardiff Bay". See text for explanation of line symbols.

According to the results obtained, "Cardiff Bay" is a region that overlaps with the areas that have been administratively labelled as Grangetown and Butetown.

The dashed lines in Fig. 6 represent three different kernel density estimations. The lightest is the one that results from points mined from Gumtree, while the darkest and most dense definition is mined from Google's user created contents. The middle gray coloured dashed contour lines stretch all over the area and are the result of delineating Cardiff Bay using the postcodes of businesses that carry "Cardiff Bay" in their business name.

With the latter representation based on business addresses, we find the most general definition of the spatial extent of "Cardiff Bay". The regions investigated during this study and represented by business data differed considerably in size compared to regions derived solely from community driven data (Gumtree, Google user created contents). Note, however, that the core of all three data sources (Gumtree, Google user created contents, and business addresses) describe an area of similar spatial extent.

Place names that exist in common usage but not in administrative geography, like the case of "Cardiff Bay," leave a number of open questions. Currently, we lack methods to validate the acquired results. Comparing models derived from different data sources is a first important step towards a representation of vernacular regions.

4.3.3 Popular Locations

A problem not discussed so far is the popularity of regions. Popular locations can skew the definition of a region towards single points (as described in Section 4.1) or to a number of densely located points. An example is given in the figure below for the "St. Fagans region" in "Cardiff", which is known for its museum. Most data points mined from the web cluster around the museum, while the rest of the region shows a very sparse point distribution. This can cause spurious peaks in the resulting kernel density estimation, as illustrated in Fig. 7. Jones, Purves et al (2008) observed the same effect when mining data from the web for "the Highlands", a region in "Scotland, UK".

To overcome this problem, another approach must be taken to calculate the kernel density estimates. Brunsdon (1995) suggests an adaptive kernel bandwidth that varies with the density of the point data. While this strategy usually smooths the kernel density surface too much, it seems to be a viable method to model popular regions with only sparse point data in the surroundings. Criteria are needed to decide when an adaptive kernel density estimation is to be favoured vs. a non parametric kernel method. Future research will address this problem.

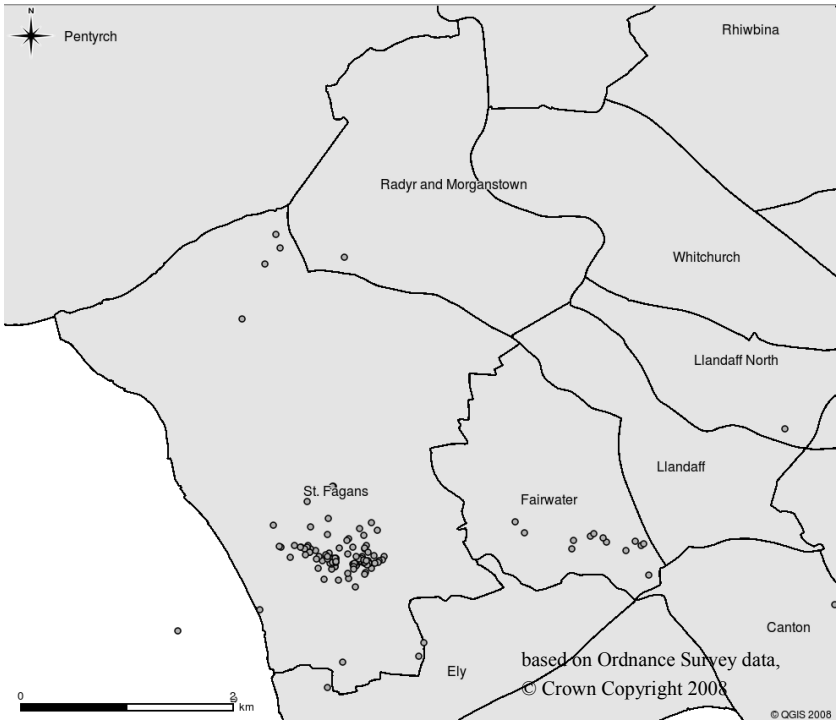

Fig. 7. Points clustering at a popular region

4.3.3 Evaluation of the Results

In the literature, administrative regions have been used to evaluate the mined data set or to train methods to detect geographic regions (Schockaert et al. 2005; Grothe and Schaab 2008). This approach is not suitable for vernacular regions, as no authority exists that defines such a region. Especially in the UK, the use of administrative definitions is open to question as a number of different definitions exist and traditionally these definitions change over time (Fairlie 1937; Association of British Counties 2008). For example, ward definitions can differ from community definitions with the same name. In Cardiff, the ward "Cathays" is split into two communities called "Cathays" and "Castle," and depending on the context, people will sometimes use the community name and sometimes the ward name. Another example is that the community "Roath" has the same spatial extent as the ward "Penylan". The evaluation of our method leaves a number of unsolved open research questions.

5. Conclusions and Future Work

The proposed method of mining user created geographic information from the Web enables the representation of the spatial extent of vernacular place names. The spatial extent of a number of places in our study region have coincided with the same named administrative boundaries. However, we have shown that this is not the case for all neighbourhoods in "Cardiff". A number of resulting representations have differed significantly from the equivalently named administrative geography.

Three different web sources have been employed to model the regions: Gumtree, business addresses, and a variety of user created geographic information sources exposed through Google's (map) search engine. A simple geocoding procedure has been utilized, extracting coordinates from web pages directly by pattern matching.

All three sources have contributed to approximating the spatial extent of vernacular place names. Representations based on user created geographic information have tended to scatter less in space than representations derived from business addresses. For all of the sources, no filtering step has been carried out. The core of the representations of the three different sources has been similar for a number of investigated regions. For some regions we could not mine enough data points from a single source.

A priority for future work is the validation of the results. Here we want to address the combination of data from different (web) sources and investigate the influence of scale. This includes experiments with different parameters for the kernel density models like determining the optimal bandwidth or a robust method to threshold the representations. We plan to carry out a web questionnaire on a large scale to gain access to an independent data source and training data for our representation method. 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 identifying a term as being a place and a method to measure the degree of vernacularity. Not all place names can be described by polygons or fields: the "Tour de France" is an event with a spatial extent that changes every year. At the same time, it is the name of a place people would look for in a search engine query. The temporal aspect of a place will have to be considered in future representations of vernacular place names.

We are especially interested in cognitive models for the representation of vernacular regions. The complexity of the problem and the variety of factors that influence human cognition of place constitute a significant challenge to future work. The results can be expected to contribute to the improvement of geographic information retrieval systems and a better understanding of people's definition of place.

7. REFERENCES

Alani, H., Jones, C. B. and Tudhope, D. (2001). Voronoi-based region approximation for geographical information retrieval with gazetteers. International Journal of Geographical Information Science **15:** 287-306.

Arampatzis, A., van Kreveld, M., Reinbacher, I., Jones, C. B., Vaid, S., Clough, P., Joho, H. and Sanderson, M. (2006). "Web Based Delineation of Imprecise Regions." Computers, Environment and Urban Systems **30**(4): 436-459.

Association of British Counties. (2008, 28th October 2007). "The problem of "county confusion" - and how to resolve it." Retrieved 17.09.2008, from http://www.abcounties.co.uk/counties/confusion.htm.

Bennett, B. (2001). Application of Supervaluation Semantics to Vaguely Defined Spatial Concepts. COSIT'01. Lecture Notes in Computer Science **2205:** 108-123.

Borges, K. A. V., Laender, A. H. F., Medeiros, C. B. and Davis Jr., C. A. (2007). Discovering Geographic Locations in Web Pages Using Urban Adresses. GIR 2007, Lisbon, Portugal, ACM, p. 31-36.

Brunsdon, C. (1995). Estimating probability surfaces for geographical point data: an adaptive kernel algorithm. Computers & Geosciences, **21:** 877-894.

Evans, A. J. and Waters, T. (2007). Mapping Vernacular Geography: Web-based GIS Tools for Capturing 'Fuzzy' or 'Vague' Entities. International Journal of Technology, Policy and Management. **7:** 134--150.

Fairlie, J. A. (1937). "Administrative Regions in Great Britain." The American Political Science Review **31**(5): 937-941.

Frank McCown and Nelson, M. L. (2007). Agreeing to Disagree: Search Engines and Their Public Interfaces. 7th ACM/IEEE-CS joint conference on Digital libraries Vancouver, ACM New York, NY, USA,p. 309-318.

Galton, A. and Duckham, M. (2006). What Is the Region Occupied by a Set of Points? 4th International Conference, GIScience 2006, Lecture Notes in Computer Science **4197**: p. 81-98.

Geiss, R. M. (2000, 2000). "Metaballs." Retrieved 17.09.2008, 2008, from http://www.geisswerks.com/ryan/BLOBS/blobs.html.

Goodchild, M. F. (2007). "Citizens as Sensors: The World of Volunteered Geography." GeoJournal **69**(4): 211-221.

Grothe, C. and Schaab, J. (2008). An Evaluation of Kernel Density Estimation and Support Vector Machines for Automated Generation of Footprints for Imprecise Regions from Geotags. International Workshop on Computational Models of Place (PLACE'08). Park City, Utah, USA.

Harpring, P. (1997). Proper Words in Proper Places: the Thesaurus of Geographical Names, MDA Information 2/3. Museum Documentation Association.

Hill, L. L., Frew, J. and Zheng, Q. (1999). Geographic Names. The implementation of a gazetteer in a georeferenced digital library, Digital Library.

Ipeirotis, P. G. A., E., Jain, P. and Gravano, L. (2006). To Search or to Crawl?: Towards a Query Optimizer for Text-Centric Tasks. SIGMOD '06: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data., ACM Press, New York, NY, USA.,p. 265-276.

Jones, C. B., Purves, R. S., Clough, P. D. and Joho, H. (2008). "Modelling Vague Places with Knowledge from the Web." International Journal of Geographic Information Systems **22**(10): 1045 - 1065

Kulik, L. (2001). A Geometric Theory of Vague Boundaries Based on Supervaluation. Conference on Spatial Information Theory - COSIT 2001, Lecture Notes in Computer Science **2205**: p. 44-59.

Mapki. (2008). Retrieved 17.09.2008, from http://mapki.com/wiki/Google_Map_Parameters.

Montello, D. R., Goodchild, M. F., Gottsegen, J. and Fohl, P. (2003). Where's downtown?: Behavioral methods for determining referents of vague spatial queries. Spatial Cognition and Computation. **3:** 185-204.

O' Sullivan, D. and Unwin, D. J. (2002). Geographic Information Analysis, Wiley.

Overell, S. E. and Rüger, S. (2007). Geographic Co-occurrence as a Tool for GIR. Workshop on Geographic Information Retrieval (GIR'07). Lisbon, Portugal,p. 71-76.

Pasley, R., Clough, P. and Sanderson, M. (2007). Geo-Tagging for Imprecise Regions of Different Sizes. Proceedings of Workshop on Geographic Information Retrieval GIR'07**:** 77-82.

Purves, R., Clough, P. and Joho, H. (2005). Identifying Imprecise Regions for Geographic Information Retrieval Using the Web. GIS Research UK 13th Annual Conference - GISRUK 2005, Glasgow,p. 313-318.

Rosch, E. (1978). Cognition and Categorization. E. Rosch and B. B. Lloyd, Lawrence Erlbaum Publishers**:** 27-48.

Schockaert, S. and Cock, M. D. (2007). Neighborhood restrictions in geographic IR. SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA, ACM Press**:** 167-174.

Schockaert, S., Cock, M. D. and Kerre, E. E. (2005). Automatic Acquisition of Fuzzy Footprints. On the Move to Meaningful Internet Systems 2005: OTM 2005 Workshops, OTM Confederated International Workshops and Posters (SeBGIS 2005). **3762:** 1077-1086.

Schockaert, S., Smart, P. D., Abdelmoty, A. and Jones, C. B. (2008). Mining Topological Relations from the Web. DEXA Workshops 2008,p. 652-656.

Silverman, B. W. (1986). Density estimation: for statistics and data analysis. Chapman and Hall, London.

Thurstain-Goodwin, M. and Unwin, D. (2000). Defining and delineating the central areas of towns for statistical monitoring using continuous surface representations. Transactions in GIS. **4:** 305-317.

Vögele, T., Schlieder, C. and Visser, U. (2003). Intuitive modelling of place name regions for spatial information retrieval. COSIT - Conference on Spatial Information Theory, Ittingen, Switzerland, Lecture Notes in Computer Science **2825**: p. 239-252.

Zelinsky, W. (1980). North America's vernacular regions. Annals of the Association of American Geographers. **70:** 1-16.