Mining the Web to Detect Place Names

Florian A. Twaroch
Cardiff University
School of Computer Science
Cardiff, UK
+442920876058
f.a.twaroch@cs.cf.ac.uk

Philip D. Smart
Cardiff University
School of Computer Science
Cardiff, UK
+442920876058
p.smart@cs.cf.ac.uk

Christopher B. Jones
Cardiff University
School of Computer Science
Cardiff, UK
+442920874796
c.b.jones@cs.cf.ac.uk

ABSTRACT
With the aim to improve the quality of gazetteers for geographic information retrieval systems, we present a method to detect place names employed by people submitting information to web resources. We investigate how often people refer to a place using locative phrases in web queries and address the problem of defining cognitively significant place names. We propose web mining as a means to decide whether a given particular named entity is in fact a place.

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

General Terms
Measurement, Experimentation, Human Factors

Keywords
Web Mining, Gazetteers, Vernacular Place Names

1. INTRODUCTION
Place names play a key role in formulating queries for geographical information retrieval, notably on the Web [3]. Gazetteers provide the main source of knowledge with which to define a footprint associated with place names in the query and in documents. 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 geographic information systems that use these gazetteers because people often use vernacular place names that are not recorded and hence result in failure to process a query (or a document) that contains such a name.

The detection of new place names is a crucial aspect of building and maintaining gazetteer services. We analyse data mined from the Web to classify names as being place names. We mine an initial candidate list of names from a social web site. For all names we generate web queries using locative phrases, counting the number of returned documents. In a first step place names found in the Ordnance Survey 50k (OS50k) gazetteer are filtered out. The document frequencies of these place names are analysed to define criteria for finding place names not yet included in the gazetteer. To measure how much a name is used by people we send web queries to selected sources and analyse the number of returned documents.

GIR'08, October 29–30, 2008, Napa Valley, California, USA.

2. STATE OF THE ART
The generation of gazetteers is a costly and labour intensive endeavour. A step towards automated building of gazetteers is the detection of place names in web documents. A rule base approach has been proposed by [1], e.g. the capitalization of a noun in an unstructured text has been used as a heuristic to detect place names and gives evidence that a named entity is a place name.

In order to insert a newly detected place name into a gazetteer it needs to be grounded, i.e. georeferenced with a set of coordinates. If web pages that include a regional vernacular place also contain references to other contained places, then coordinates for these other place names may be found by conventional gazetteer lookup hence providing a set of locations approximating the extent of the vernacular place [3]. Volunteered geographic information and social web sources are another means to ground newly detected names [2].

We focus here on the detection of place names in web documents and do not discuss the process of generating the associated footprint. Trigger phrases based on English spatial prepositions combined with Web counts serve to identify place names. Web counts have been previously used to measure the the cognitive significance of landmarks [4].

3. DETECTING PLACE NAMES
With web scraping software 2500 distinct location entries were mined from a social web source (Gumtree - http://www.gumtree.com) for the region of Cardiff, UK. The data represents location tags freely entered by people to sell/buy items or make social contacts. Although users are advised to enter 'suburb only' locations, the mined tags can contain post codes, telephone numbers, names of roads, streets, etc. and other information, e.g. further promoting the item.

We first filtered out tags that contained addresses (227 streets/places/avenues) or numbers (115 numbers) using regular expressions. Then we created trigger phrases using the following spatial prepositions: in, inside, within, at, near, around, across, nearby, out of, toward, through, from, to, over, close to, off the north|south|west|east of and via to classify names as place names. We then counted the number of documents returned to web

queries using Yahoo's BOSS API. For each of the candidate names we could evaluate counts of locative property phrases.

Web document counts can be raised misleadingly by ambiguities and their consideration is a prerequisite towards a successful detection of new place names. Therefore we calculated the ratio of counts received through web queries containing the newly detected place name vs. the web counts of queries containing the place name and the region "Cardiff". This ratio is close to 1.0 for terms that solely appear on web pages together with Cardiff and falls towards 0.0 for terms that are also related to other concepts or locations, i.e. they are ambiguous. As our initial list of candidate names were just locations in the vicinity of Cardiff we filtered out names whose web count ratio was below 0.75.

Table 1. Document frequencies for gazetteer names

Relation	Min	Median	Max
at	0	57699	1750000
in	0	19500	11600000
near	0	8135	197000
...	...	...	...
combined	15	27620	15050000

The initial list mined from Gumtree contained 224 tags that were also found in the OS50k gazetteer. Table 1 shows for selected locative phrases that the number of document hits can vary between 0 and more than 11 million. We tested 20 different spatial relationships and found that there is no general preference for a certain spatial relationship in connection with a place name from the gazetteer. Consequently to create thresholds to classify names that are not yet in the gazetteer a combined measure of different spatial prepositions has to be used. Combining all the spatial prepositions in web queries shows that for any of the given names in the OS50k gazetteer at least 15 hits can be expected.

Although people are asked to freely enter a place name into the location tag in Gumtree (labelled 'suburb only') only roughly 10% of the entered names will be found in the OS50k Gazetteer. At this point we did not investigate how many misspellings occurred in the data set.

4. RESULTS

In order to detect new names we measure the variation across different spatial prepositions for a given name from the candidate list. Thresholds are set for a minimal amount of 15 returned web pages. One wants to be careful when adding new names to a gazetteer and prefer to accept a high number of false negatives than false positives. Therefore we decided to use very conservative thresholds. We identified the following names as candidate names to be added to the gazetteer:

Atlantic Wharf, Cardiff Gate Business Park, Cardiff Cardiff Bay, Cardiff and Vale of Glamorgan, Culverhouse Cross, Cardiff East, pontprennau; [..] millenium stadium, [..] Cardiff International Arena, cathays terrace, Cardiff Gate, Wales Millennium Centre., [..], Rhoose, Cardiff North, Cardiff University, [..], Cardiff City Centre, South Glamorgan, Heath Hospital, Penarth Marina, Cardiff Centre, East Canton, Century Wharf, Cardiff Central

Note that some of the names found are not capitalized. A major drawback of the present method is the simplified analysis of text on web pages through search engines which are for example very generous in the use of stop words. Proper text mining will help to identify more place names and avoid false negatives. In order to generate more accurate counts a focused crawling of web documents is necessary. We are currently building such a collection of web documents.

False negatives occur mainly because of not considering geo-ambiguities. *Lakeside* is a name in the gazetteer that came up in our candidate list but was not found after the naïve filtering step. False positives occur because we did not consider misspelling of names and abbreviations require a further normalization step. Some of the location tags contain place name hierarchies that have not been identified by our method. The performance of the method is also due to using a list of candidate names that have been tagged as locations.

5. CONCLUSIONS

Even for place names found in the gazetteer the number of different associated spatial prepositions mined from the Web is very small. Semantic knowledge is necessary to detect place names based on document counts and a concept ontology may help to resolve ambiguities caused by concepts. Co-occurence measures appears beneficial to identify geo ambiguous place names [4].

Future work will address a number of open questions such as 1) the use of place names across different web sources 2) other criteria and trigger phrases to identify place names 3) combination of different criteria 4) detection of place names with other/no inital name lists. The integration of newly detected place names into an existing gazetteer leaves still plenty of open questions.

6. ACKNOWLEDGEMENTS

We gratefully acknowledge 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] Maynard D., Bontcheva K. & H. Cunningham. Automatic Language-Independent Induction of Gazetteer Lists, LREC 2004, http://gate.ac.uk/sale/lrec2004/gazcollector.pdf

[2] Popescu, A.: Greffenstette G., Moellic P.: Gazetiki: Automatic Creation of a Geographical Gazetteer, JCDL 2008

[3] Purves, R., P. Clough, and H. Joho. Identifying imprecise regions for geographic information retrieval using the Web. GISRUK2005.

[4] Tezuka, T.; Yokota, Y.; Iwaihara, M. & Tanaka, K. Extraction of Cognitivley-Significant Place Names and Regions from Web-based Physical Proximity Co-occurences, WISE 2004, LNCS 3306, Springer, 113-124