Modelling Confidence in Extraction of Place Tags from Flickr

Omair Z Chaudhry[1], William A Mackaness[2]

[1] Manchester Metropolitan University, Chester Street, Manchester
Tel.+44 (0) 161 247 1574
Email: O.Chaudhry@mmu.ac.uk
[2]University of Edinburgh, Drummond Street, Edinburgh

Summary: The volume and potential value of user generated content is ever growing. One such valuable source for better understanding of naïve or vernacular geography is in the form of geotagged images from Flickr. Research in the past has looked into automatic identification of place tags from this source. This paper gives an overview of a data mining techniques for identification of places tags at different levels of detail and a Bayesian inference model to predict the probability for each selected tag as being 'non-noise'.

KEYWORDS: data mining, visualisation, place tags, bayesian inference, Flickr

1. Introduction

There is increasing interest in mining the 'geography' that is now stored on the web. The 'Geospatial web' affords a capacity 1) to search for documents and imagery based on references to the geography (Hill et al. 2000); 2) to model vernacular geographies (Hollenstein and Purves 2010; Jones et al. 2008; Lüscher and Weibel 2010); and 3) to support more intuitive use of web mapping technologies. More broadly it enables us to think differently about how we do GI science (Kuhn 2007). Research on the Geospatial web is fuelled by freely available user generated content (UGC) or Volunteered Geographic Information (VGI) (Goodchild 2007). Open Street Maps, Wikimapia, WikiLocation, Geonames are frequently cited examples of VGI, and in some contexts rival conventional ways of capturing geographic information (Howe 2008). However, the very nature of UGC means that it is often inconsistent, incomplete and poorly structured (Purves 2011). Tags attached to images and videos on data sharing services such as Flickr, and YouTube may contain any number of references to places, objects and events but not in a form that can be readily understood except by people with some knowledge of the vocabulary used. For the tags associated with picture in Figure 1, we might ask: is this Edinburgh UK, or Edinburgh US? Is New Town an area in Edinburgh? Is Nikon a camera, place or name of person? In other words, how might we extract the 'meaningful' information among the tags used to describe this image, and how might it be structured so as to facilitates its retrieval and use?

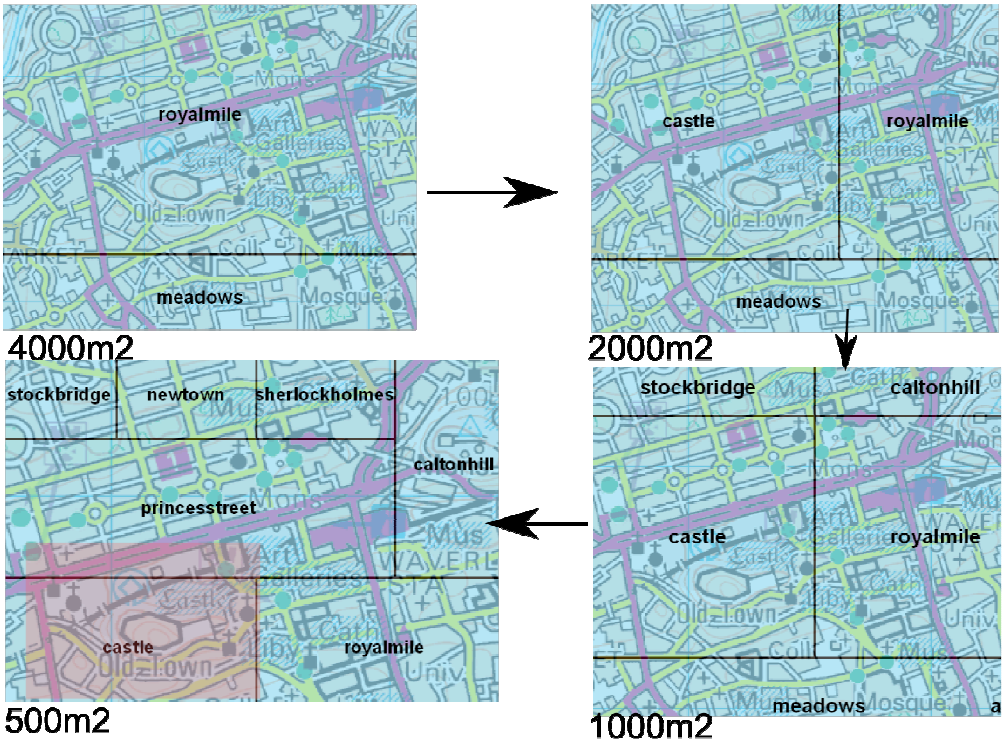
Figure 1: A geotagged image with tags: Arthurs seat, New Town, Nikon, John, Edinburgh.

1.2 Place Tags from Flickr

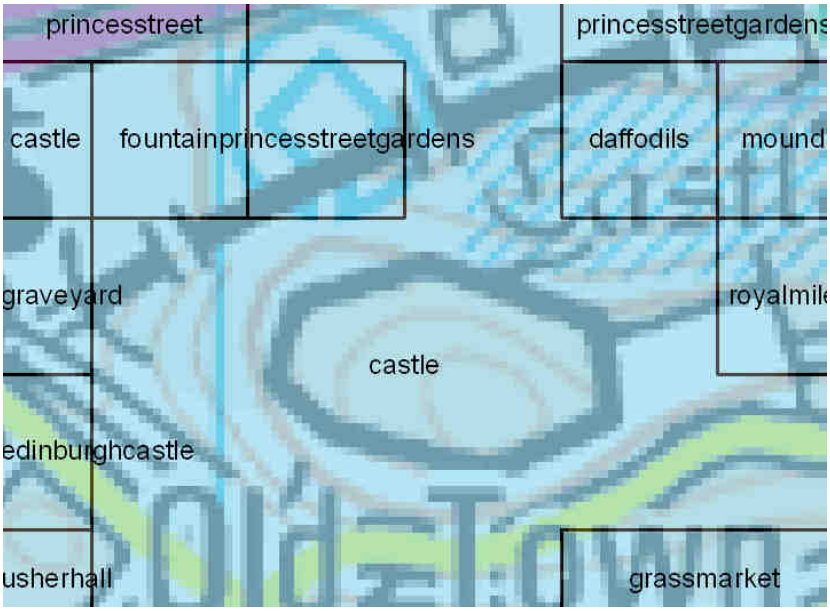
Flickr is one the biggest sources of images on the web (Flickr 2011). There are estimated to be 5.9 billion pictures available on Flickr. The Flickr website (http://www.flickr.com/map/) suggests that as of July 2011 there are more than 153 million pictures that have been geo-tagged (images that have been assigned a location). The unstructured nature of Flickr tags is both its strength and weakness. An image can be freely tagged with any number of tags, using any words or phrases – thus affording maximum flexibility. The weakness lies in absence of categories and structure – with the challenge of identifying patterns and meaning among the noise of the tag descriptors. In response to this, various researchers have presented techniques for extracting structured information from this data (Rattenbury and Naaman 2009; Jaffe et al. 2006; Ahern et al. 2007; Girardin et al. 2008).

Most of the techniques presented in research, for extraction of place tags from VGI dataset such as Flickr, are usually based on principles of information retrieval, data mining, and web harvesting. Once meaningful tags have been selected by such techniques they are usually evaluated manually or compared against manually selected list of tags for the given area (Rattenbury and Naaman 2009). Our previous research (Chaudhry and Mackaness 2012) presented a similar approach based on TF-IDF (Purves 2011) for selection of a place tag for each grid cell – varying the size from $100m^2$ to $4000m^2$ for the city of Edinburgh (Figure 1).

Once selected we used our local knowledge of Edinburgh to manually inspect the tags to assess the veracity of the tag. Was it 'noisy' (Elaine's Wedding, Nikon) or not (Paris, Grand River). In the context of Edinburgh, *manual* inspection of the selected tags at $100m^2$ revealed that out of a total of 3,951 cells (at $100m^2$) that were assigned a tag, only 34% were assigned a meaningful place name; the remainder were deemed 'noise'. Similar manual inspections were carried out at all the grid resolutions (Figure 2 and 3).



(a)



(b)

Figure 2: (a) Selected labels from tags at different levels (4000m2 to 500m2) of spatial detail. (b) Selected tags at 100m2 for region highlighted in (a) at 500m2

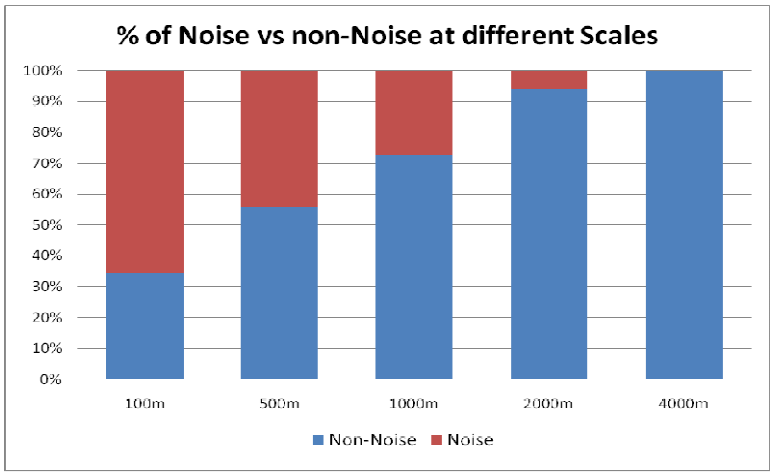


Figure 3: Result of manual inspection to differentiate between noise and non-noise tags selected using the above approach for Edinburgh, UK

2. Post Selection Refinement using Data mining Techniques

In this research we present a data mining technique (using Bayesian Inference) to build a model using a number of (independent) variables computed for each selected tag from the source data (Flickr dataset) to generate a confidence value, representing the probability that the selected tag is indeed *not noise*, attached to each selected tag. We used the manual classification to build and test the accuracy of the approach. We randomly selected 70 % of the selected tags at $100m^2$ to build the model. The remaining 30% of the selected tags at $100m^2$ and all the selected tags for the rest of the scales ($500m^2$, $1000m^2$, $2000m^2$ and $4000m^2$) were used to assess the accuracy of the model.

2.1 Bayesian Inference

Bayes' Rule is a simple way of calculating conditional probabilities (Hacking 2001). Conditional probabilities are those probabilities whose value depends on the value of another probability (Duda et

al. 2001). The Bayesian decision rule tries to minimize the probability of error in a decision by deciding the most probable outcome.

Using a Bayesian approach, we can answer questions of the following form: 'For a given, selected tag for a cell with a specific set of characteristics, what is the likelihood that it belongs to the list of non-noise tags with their specific set of characteristics?'. What is returned is a probability value reflecting the likelihood that the tag should indeed be selected as the tag for that cell. Here we used an approach similar to that proposed by Lüscher et al (2009). They used normal kernel density estimation techniques to determine joint probability density values for each building feature in the Topography Layer of Ordnance Survey MasterMap. In their research, the value was then used to classify buildings into terraced or non terraced houses. In essence this involved comparing an unknown with a sample of 'knowns' (training sample) and classifying the unknown according to how similar it was to the 'knowns'.

The joint conditional probability for a classification of an unknown is given by Equation 1.

$$P_c(\vec{f} \mid C = c) = \frac{1}{N \parallel \vec{h} \parallel} \sum_{i=1}^{N} K\left(\frac{\vec{f} - \vec{f}_i}{\vec{h}}\right) \quad (1)$$

Where Pc is the conditional probability of the unknown for the predicted class C, N is the number of samples in the training dataset, $\vec{h}$ are the bandwidths, K is the standard normal distribution function, $\vec{f}$ is the vector of properties of unknown, $\vec{f}_i$ is the vector of the same properties of training dataset.

The 70% of randomly selected tags were used as the training dataset and the remaining 30% at 100m^2 and the rest of the dataset at lower levels of detail (500m2-4000m2) were 'withheld' in order to test the accuracy of the resultant model. Table 1 presents the result of the Bayesian approach for all the unselected cases i.e. the non-training dataset for Edinburgh at all levels of detail. Three attributes (or properties) were used for Bayesian classification namely: the unique user frequency of the selected tag within the cell; unique user frequency of the selected tag in the whole collection; and the selected tag frequency in the whole collection. In Table 1 'Class 0' represents those tags that are deemed to be noise and 'Class 1' represents tags that are non-noise (meaningful). The overall accuracy of classification of the non-training dataset across all levels using this model was above 80% (Table 1). This means that the model is more than 80% accurate in being able to distinguish between noise and non noise tags.

Table 1: Classification result by Bayesian inference for Edinburgh at different levels of detail

		Predicted Class			
		Class	**0**	**1**	**% Correct**
Scale: 100m2	Original Class	0	727	61	92.26
		1	113	282	71.39
				Overall %	**85.29**
Scale: 500m2	Original Class	0	239	8	96.76
		1	72	245	77.29
				Overall %	**85.82**
Scale: 1000m2	Original Class	0	48	3	94.12
		1	30	101	77.10
				Overall %	**81.87**
Scale: 2000m2	Original Class	0	2	2	50.00
		1	7	39	84.78
				Overall %	**82.00**
Scale: 4000m2	Original Class	0			
		1		10	100.00
				Overall %	**100.00**

We linked the probability values calculated by the model to each tag and created an interactive tree view visualization in order to visualise these hierarchal relationships between tags and also to explore the probability values in more detail. The tags are connected hierarchically via their spatial relationship – 'contained by' (Figure 4). The number next to each tag name in Figure 4 show how confident (the probability value) the model is that the tag is *not noise* at that particular granularity. The visualization is also available as an applet http://omairchaudhry.net84.net/City_Viz/Tree_View_Tags_Confidence_Bayes.html

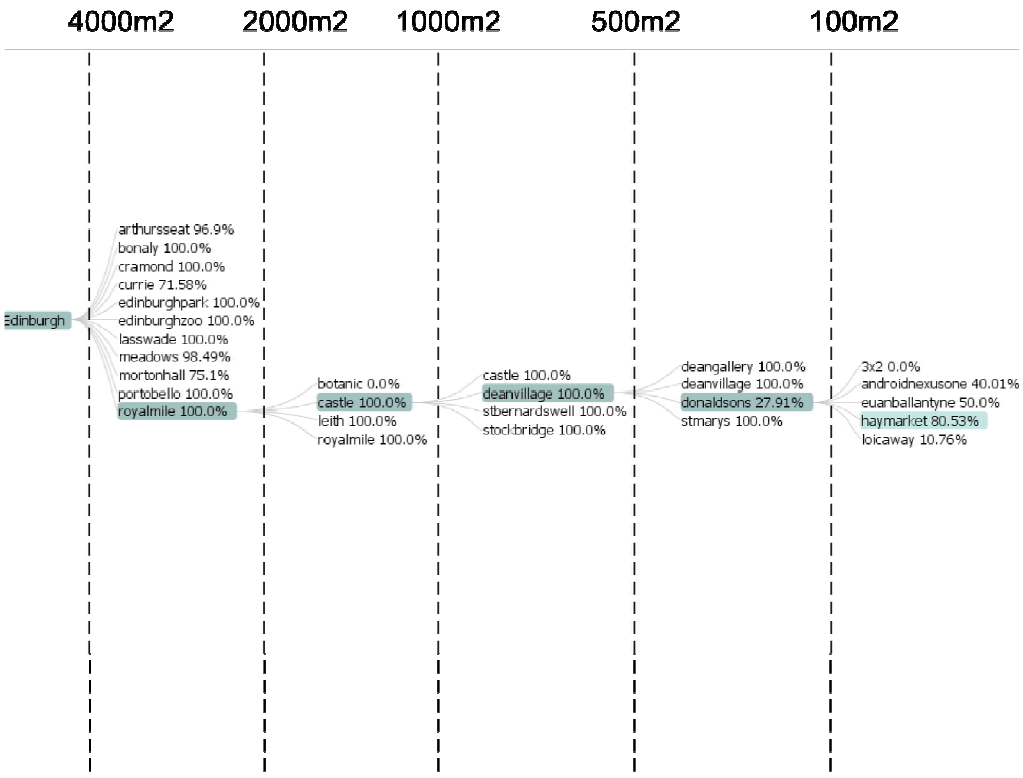


Figure 4: Tree view visualisation of selected tags and their confidence (probabilty) value as predicated by the model

3. Conclusion

The geo-tagged images generated by the public and freely accessible via a number of Web2.0 services such as Flickr offer great potential to understand people's perception of places and points of interest. A lot of research in the Geospatial web has explored the use of Flickr tags as a source for vernacular geography and in automatic selection of meaningful place tags from this source but there has been limited research in exploration of the reliability of these tags at different scales and also in the development of models that can calculate the probability of a tag being 'meaningful'. In this research we have presented one such model and demonstrated the accuracy of the model at various levels of detail. From our case study of city of Edinburgh the model was tested at five different levels of detail and the overall accuracy of the model was at least 80% at each level of detail.

4. Biography

Omair Z Chaudhry is a lecturer in GIS at Manchester Metropolitan university. His research interests are in geo-visualisation, spatial data mining and map generalisation.

William A Mackaness is a senior lecturer at the University of Edinburgh. His research interests are in spatial data abstraction, visualisation and location based services.

References

Ahern, S., Naaman, M., Nair, R. and Yang, J. (2007) World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections *Seventh ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) ACM*, pp. 1-10. New York.

Chaudhry, O.Z. and Mackaness, W.A. (2012) Automated Extraction and Geographical Structuring of Flickr Tags. *GeoProcessing* Valencia, Spain, 30 Jan - 4 Feb

Duda, R.O., Hart, P.E. and Stork, D.G. (2001) *Pattern classification.* John Wiley & Sons, New York.

Girardin, F., Calabrese, F., Fiore, F.D., Ratti, C. and Blat, J. (2008) Digital Footprinting: Uncovering

Tourists with User-Generated Cotent. *Pervasive Computing* 7, 36-43.

Goodchild, M.F. (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69, 211.

Hacking, I. (2001) *An introduction to probability and deductive logic*. Cambridge University Press, Cambridge.

Hill, L.L., Carver, L., Larsgaard, M., Dolin, R., Smith, T.R., Frew, J. and Rae, M.-A. (2000) Alexandria digital library: user evaluation studies and system design. *Journal of the American Society for Information Science*, 51, 246-259.

Hollenstein, L. and Purves, R.S. (2010) Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 1, 21-48.

Howe, J. (2008) *Crowdsourcing: How the Power of the Crowd is Driving Business.* 1 Crown Business.

Jaffe, A., Naaman, M., Tassa, T. and Davis, M. (2006) Generating summaries and visualization for large collections of geo-referenced photographs. *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. ACM, Santa Barbara, California, USA.

Jones, C.B., Purves, R.S., Clough, P.D. and Joho, H. (2008) Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, 22, 1045-1065.

Kuhn, W. (2007) Volunteered Geographic Information and Geographic Information Science. *NCGIA and Vespucci Specialist Meeting on Volunteered Geographic Information* (ed M. F. Goodchild). Santa Barbara, CA.

Lüscher, P. and Weibel, R. (2010) Semantics Matters: Cognitively Plausible Delineation of City Centres from Point of Interest Data. *13th workshop of the ICA commission on Generalisation and Multiple Representation*. Zurich, Switzerland.

Lüscher, P., Weibel, R. and Burghardt, D. (2009) Integrating ontological modelling and Bayesian inference for pattern classification in topographic vector data. *Computers, Environment and Urban Systems*, 33, 363-374.

Purves, R.S. (2011) Methods, Examples and Pitfalls in the Exploitation of the Geospatial Web. *The Handbook of Emergent Technologies in Social Research* (ed S. N. Hesse-Biber), pp. 592 -624. Oxford University Press, Oxford.

Rattenbury, T. and Naaman, M. (2009) Methods for extracting place semantics from Flickr tags. *ACM Trans. Web*, 3, 1-30.