Automated Extraction and Geographical Structuring of Flickr Tags

Omair Z Chaudhry[1], William A Mackaness[2]
[1)]Manchester Metropolitan University, Manchester UK
[2)]University of Edinburgh, Edinburgh, UK,
Emails: O.Chaudhry@mmu.ac.uk, william.mackaness@ed.ac.uk

Abstract— **The volume and potential value of user generated content (UGC) is ever growing. One such source is geotagged images on Flickr. Typically, images on Flickr are tagged with location and attribute information variously describing location, events or objects in the image. Though inconsistent and 'noisy', the terms can reflect concepts at a range of geographic scales. From a spatial data integration perspective, the information relating to 'place' is of primary interest and the challenge is in selecting the most appropriate tag(s) that best describe the geography of the image. This paper presents a methodology for searching among the 'tag noise' in order to identify the most appropriate tags across a range of scales, by varying the size of the sampling area within which Flickr imagery falls. This is applied in the context of urban environments. Empirical analysis was then used to assess the correctness of the chosen tags (whether the tag correctly described the geographic region in which the image was taken). Logistic regression was then used to build a model that could assign a probability or confidence value to each selected tag as being a appropriate geographic tag. The high correlation values achieved bodes well for automated environments - environments in which this methodology could be used to automatically select meaningful tags and hierarchically structure UGC in order that it can be semantically integrated with other data sources.**

Keywords-data mining; information retrieval; vernacular geography; granularity modelling.

I. INTRODUCTION

The geospatial web comprises multiply sourced data (both formal and unstructured). Formal geographies (provided by National Mapping Agencies (NMA) and Government Bodies) reflect an *administrative* view of geography. Whereas User Generated Content (UGC) reflects observations at different levels of detail, more qualitative in nature, and relating to ideas of 'place' (events and performance) rather than formal and systematic descriptions of space. These two types of data offer complementary and synergistic approaches to the mining and intuitive understanding of geographic information. The conflation of 'formal' and 'informal' (such as free access to NMA datasets via Open Street Map) reflects a blurring of this binary but data integration is far more than simple overlay. Much has been written on the need for semantic and ontological modeling in order to automatically conflate the qualitative and the quantitative [1, 2]. The difficulty being the vagueness omnipresent in the geospatial domain, the problematic notion of space and place, and the granularity inherent in the description of geographic concepts [3].

There is increasing interest in mining the 'geography' now stored on the web. Geography provides a context and an intuitive way of organizing digital information. The 'Geospatial web' reflects a capacity to search for documents based on references to the geographical (using geotags [4]), to model vernacular geographies [5-7] and to support web mapping technologies [8]. Research on the Geospatial web is fuelled by freely available user generated content (UGC) or Volunteered Geographic Information (VGI) [9]. Open Street Maps, Wikimapia, WikiLocation, Geonames are frequently cited examples of VGI, and in some contexts rival conventional ways of capturing geographic information [10]. But the very nature of UGC means that it is often inconsistent, incomplete and poorly structured [11]. Tags attached to images and videos on data sharing services such as Flickr, and YouTube may contain a 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 example, for Fig 1, how might we extract the 'meaningful' information inherent in the images and tags, and how might we structure the geography implicit in this image in a way that facilitates its retrieval and use. Is the tag 'Paris' in Fig. 1 the name of a person, a world capital or a community in Ontario, Canada? This is example of non-geo/geo ambiguity similarly there can be geo/geo ambiguity for example 'London' in UK or 'London' in Canada. A number of techniques have been proposed to 1) automatically unambiguously extract place names, and 2) assign them spatial coordinates [11]. These two process are commonly referred to as geo-parsing and geo-coding respectively [12]. This paper describes a technique for automatically retrieving and visualizing 'meaningful' place names from a VGI dataset (specifically Flickr geo tagged images) at different spatial levels of detail.

Section 2 describes a methodology to mine information from a list of geotags and to sample data at different granularities in order to hierarchically structure the data. Section 3 uses data mining techniques for 'post selection' of the tags that seeks to filter out selected tags that are not geographical in nature. The section also presents

 ISBN: 978-1-61208-178-6

visualization techniques used to convey the hierarchical structure and confidence in the selection of tags.

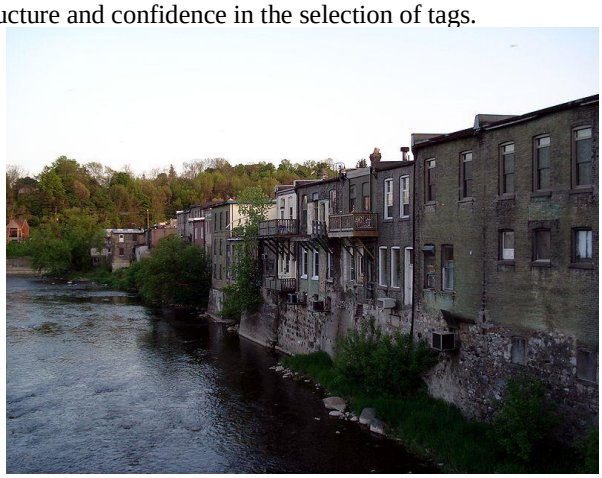

Figure 1. Geotags: Paris, Grand River, Nikon: Is Paris a person, a world capital, or a community in Ontario,Canada? Was Nikon the person who took the picture?

II. FLICKR IMAGES AND THEIR TAGS

The aim is the extraction of meaningful place names and points of interest from among the tags that people associate with the Flickr images they take. Flickr is one of the biggest sources of images on the web. There are estimated to be 5.9 billion pictures available on Flickr. The Flickr website [13] suggests there are more than 153 million pictures that have been geo-tagged - pictures that have been assigned a geographic footprint (a latitude-longitude coordinate). Users are free to assign any number of tags to a picture. The tags can be city names, landscape descriptions, events, the camera used, dates, regions within a city, gardens, places and features of interest, indeed any adjective you care to imagine!

Various authors have presented techniques for extracting structured information from data [14-17]. An assumption common to these research efforts, is that the image tags variously connect the image with a particular geography of place and space (idea of 'place semantics' [14]). Among other things, the truth of such an assumption can be corroborated against the density of nearby images and diversity of image takers. The value in extracting such place semantics are well understood (e.g., improved search, intuitive (vernacular) descriptions of space, automated assignment of place semantics to untagged imagery). Most of the research has focused on extraction meaningful tags on the same level of spatial detail. Also such research uses manual comparison for testing accuracy of the approach. Here this research focuses on extraction of meaningful tags at *different* levels of detail and automatic assignment of confidence value (probability of correctness) by a model.

A. *Accessing Flickr*

Flickr provides a non-commercial API in order to access its dataset. The API provides a number of ways in which the Flickr images can be queried: by date, tags, geographic location, or groups for example. In addition there are a number of free Flickr API programming kits available. These kits are programming interfaces for different programming languages (notably C, Java, and Python [24]). These kits allow API queries to be embedded within user's own code. For this research we used flickrj – a java Flickr API kit which was used to extract the Flickr dataset for the City of Edinburgh, Scotland.

In order to obtain all publicly geo-tagged images for the city of Edinburgh we overlaid a matrix of regular cells, each of 100m2 covering the whole city. A total of 134,986 images with its id, user tags, URL, user id, latitude, longitude values were thus obtained for the whole city of Edinburgh. There were a total of 20,400 100m2 cells covering the city of Edinburgh. Only 3,993 of those cells contained one or more Flickr image that were tagged (Fig. 2).

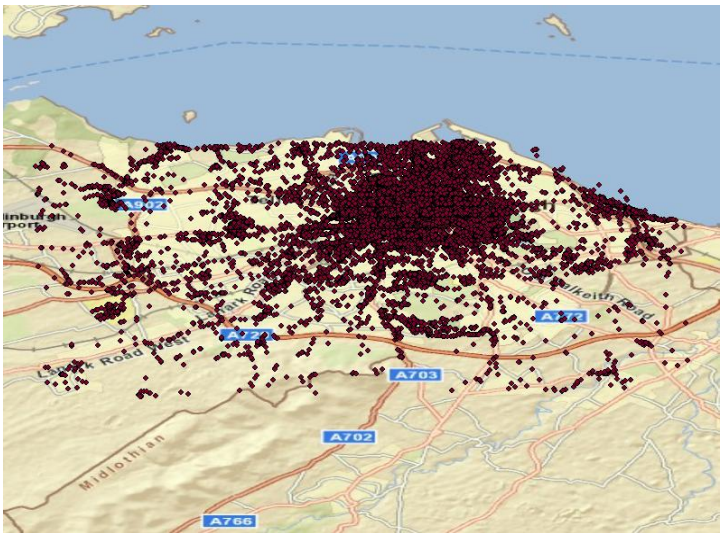

Figure 2. A map of Edinburgh showing the distribution pattern of Flickr images across the city.It shows the distribution is not homgenous across the whole city

B. *Selection of 'meaningful' place names for each Cell*

The next step was to assign to each cell, a place name or point of interest chosen from the associated image tags falling in that cell. For each cell we selected all the images and their respective tags. From these tags we then selected the most 'appropriate' tag for that cell. The simplest text analysis technique is to rank the tags according to frequency of occurrence and select the most frequently occurring. Unsurprisingly for many of the cells, the most frequently occurring tag was either 'Edinburgh' or 'Scotland' – not a tag that reflects the 'local' spatial granularity of a 100 m^2 cell! The second problem is one of 'tag distortion' arising from a single person, taking a relatively large number of images, and using the same tag to describe an event (rather than a place). For example one cell was named: 'Elaine's wedding'. This was because the tag was associated with 20 separate images spatially contained by that particular cell.

 ISBN: 978-1-61208-178-6

1) Modeling the Local Context

In order to incorporate the local context, we applied the TF-IDF algorithm (term frequency – inverse document frequency). TF-IDF is a well know information retrieval technique and is used to weight the importance of an individual term's contribution in relation to relevance in a search [11]. In our study for each tag contained by a cell we computed its term frequency (TF) by dividing the number of times the tag occurred within the cell by the total number of all tags within that cell. Inverse document frequency (IDF) was computed by taking the logarithm of the total number of cells that contain any tag (i.e. 3,993) divided by the total number of cells that contain that particular tag. TF-IDF is the product of TF and IDF. This product (TF-IDF) is the resultant weight assigned to the tag. The highest weighted tag is then selected for each cell. This approach ensures that those tags which are frequently inside one cell but occur rarely in other cells are given a high weight. So 'Princes Street' (a major high street in the city) will have a high weight, and 'Edinburgh' or 'Scotland' will have a low weight since they occur frequently within the cell as well as in the whole collection.

2) Object View vs Subject View (User Frequency + TF-IDF)

The TF-IDF approach identifies tags 'local' to a region, but it does not remedy the problem of 'tag distortion' (the example of 'Elaine's Wedding'). We can resolve (to large extent) this problem if we take an object's perspective of the tag, rather than a subject's perspective. We might realistically expect different people to use the same tag, thus corroborating the validity of that tag. In the example of 'Elaine's Wedding', it is very unlikely that other people would use this tag in the same cell. So by attaching importance to the number of different users who use a particular tag (the idea of collective intelligence), we might overcome the distorting effect of a single user attaching the same tag to multiple images falling in the same cell. So instead of using tag occurrences we use user frequencies associated with each tag in order to calculate TF-IDF weights. Using the user count reduces the TF for tags such as 'Elaine's wedding' and IDF ensures that tags with a high user count, such as 'Edinburgh' and 'Scotland', will have low IDF values. This results in low weights (TF-IDF) for both of these types of tags. All tags contained by a particular cell are then sorted in descending order and the tag with the highest weight is selected. There is a proviso: the tag is selected only if it has a 'user count' of at least two. The extra condition ensures that at least two distinct users have used the same tag. If this condition is not met then the next tag in the sorted list is checked and so on until both conditions are met. This process resulted in 3,951 cells being assigned a tag at the 100m2 spatial resolution for the city of Edinburgh.

C. Hierarchical Structuring and Visualisation

We can imagine people's understanding of the city to be hierarchical in nature (Fig 3), comprising high streets, shopping centers, and business districts, at one level, suburbs, districts, parks at another, all partonomically constituting the city [18].

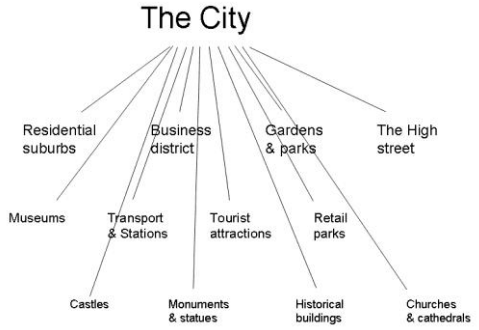

Figure 3. An example of conceptual view a city

Therefore, as a next step, we applied the same methodology but to increasingly larger cell sizes, covering the same region, in order to try and mirror a linked hierarchical structure. We applied the technique to grid cells with resolution: 500m2, 1000m2, 2000m2, 4000m2 covering the city of Edinburgh. Although the choice of scales (100m2 to 4000m2) is arbitrary but the presented approach for tag selection is applicable at any selected scale or shape and size of grid cell. Once the labels were selected for each cell at a specific level (100m2 to 4000m2) we aggregated adjacent cells if they had the same label. This created regions that shared the same tag. Fig 4 shows the result of applying this approach to the city centre of Edinburgh (The Castle and The Royal Mile). Fig. 4 also shows the selected tags as labels for each cell at different levels of detail. At each higher level the most dominant tag (highest TF-IDF weight) is selected as the label.

Upon inspection of the selected tags it was apparent that there was still 'noise' present among selected tags, most notably at the finest level of detail (100m2). Date tags are an example of such noise (for example '2007'). This happens because a tag such as '2007', will have less weight only if very few distinct users have used that tag or that it is common to a whole collection of cells. It is still possible that such tags have been used by a number of distinct users within the cell. Upon manual inspection of the selected tags at $100m^2$ it was found that out of a total of 3,951 cells ($100m^2$) assigned a label, only 34% contained a meaningful place name or point of interest; the remainder was 'noise'. Similar manual inspections were carried out at all scales. As illustrated in Fig 5, the noise tags selected by this approach are highest at the most detailed level ($100m^2$). But, at lower levels of detail the TF-IDF weights of noise tags will be less as compared to non-noise tags because the spatial extent is larger and thus there are more images that have appropriate non-noise tags.

 ISBN: 978-1-61208-178-6

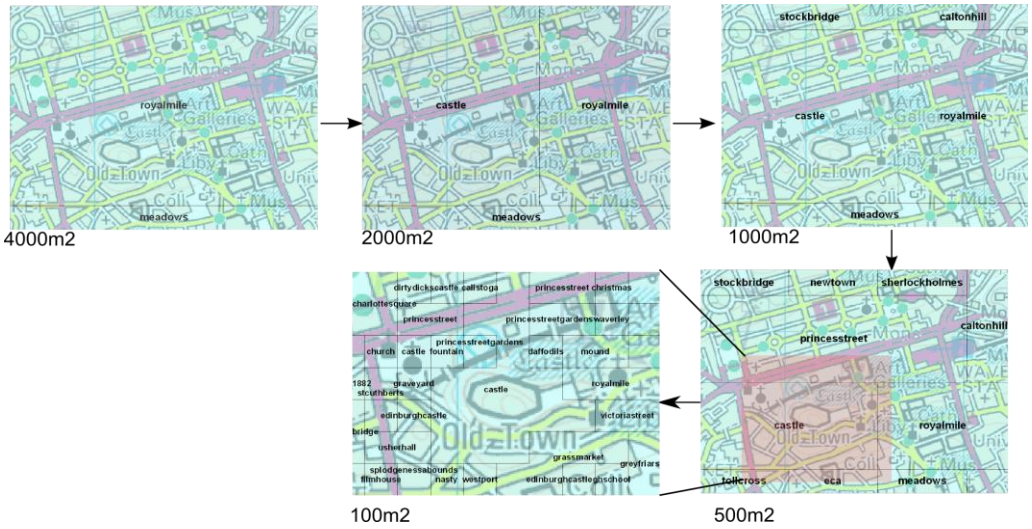

Figure 4. A conceptual view a city? Different tags at different levels of detail.

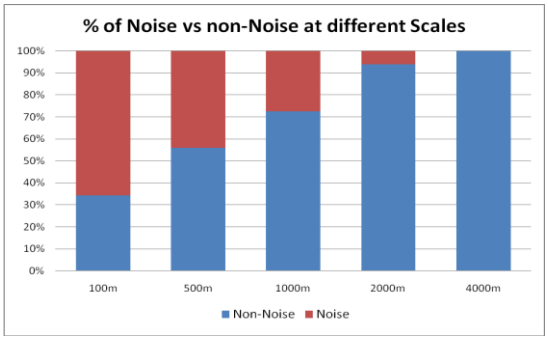

Figure 5. High amounts of noise at smaller cell size (Result of manual inspection)

III. POST SELECTION REFINEMENT

In response to the problem of 'noise' (especially at the finest level of granularity), a data mining technique was explored in order to automatically filter out this noise. In the past, techniques such as 'stop words' and 'controlled vocabularies' have been proposed to remove such noise [19, 20]. In this research, we tested a data mining technique (logistic regression) in order to build a model using a number of (independent) variables computed for each selected tag from the source data (Flickr dataset) without utilizing any other external data. In essence the aim was to further refine the above approach such that a confidence value, representing the probability that it is not noise, can be attached to each selected tag. We used a manual classification to build and test the accuracy of the approach. We randomly selected 70 % of the manually classified cases at 100m2 to build the model. The remaining 30% of the manually classified data was used to assess the validity of the approach.

A. *(Binary) Logistic Regression*

Logistic regression is similar to multiple regression except that the dependent variable in the logistic regression is sampled as a binary variable i.e. noise (y=0) or non-noise (y=1). Logistic regression therefore models the probability of presence and absence for a given observed value among the predictor variables. The probability function can be written as: [21].

$$P(Y=1) = \frac{1}{1+e^{-(\alpha+\beta_1 x_1+\beta_2 x_2+....+\beta_n x_n)}} \quad (1)$$

In Equation 1, y is the dependent variable, α is the intercept, β is the coefficient(s) of the independent variable(s) x. Equation 1 can be used to calculate the probability that the outcome (dependent variable) will be 1. In this research y is 1 if the selected tag is considered to be the correct label for a given cell, otherwise it is 0.

For each cell and its selected tag, we calculated a number of variables, (x_n). These included: the user frequency of a selected tag within the cell (x1); the user frequency of the selected tag in the whole collection (all cells) (x2); the selected tag frequency within the cell (x3); the selected tag frequency in the whole collection (x4); the total number of images contained by the cell (x5); the total user frequency for all the tags contained by the cell (x6); and the total raw frequency of all the tags contained by the cell (x7). Stepwise binary logistic regression was carried out in SPSS [22], randomly selecting 70% of the manually classified cases – the remaining 30% were used to the test the accuracy of the model.

Table I lists the selected variables (x1, x2 and x4) from the last stage of the stepwise logistic regression together with their coefficient values. Nagelkerke's R2 value for the model is 0.423. Table II lists the classification result after the final step of stepwise logistic regression. Table II shows the percentage of correctly identified cases from the 70% sample

dataset is 81.1%, and the percentage of correct results for the remaining 30% of the sample dataset is 82.3%. The cut off value used in Table II to separate between cases classified as 0 or 1 is 0.5. This simply means that if the resultant probability for a tag is 0.51 it will belong to class 1 (non-noise) and if 0.49 it will belong to class 0 (noise).

TABLE I. SELECTED VARIABLES IN THE MODEL

Variables	B	S.E.	Wald	df	Sig.	Exp(B)
x1	1.413	.090	246.095	1	.000	4.107
x2	.013	.002	46.500	1	.000	1.013
x4	-.001	.000	35.958	1	.000	.999
Intercept (α)	-2.738	.121	512.187	1	.000	.065

Once the model was built (Table I) for the most detailed scale ($100m^2$), we applied this model at all the remaining scales. The result from the model was evaluated against the manual inspection carried out in the previous section (Fig. 5) and is presented in Table III. As the scale reduces, the capacity of the model to correctly predict the result is significantly increased (Table III), especially for the true positive cases.

We linked the probability values calculated by the model to each tag and created an interactive tree view visualization in order to explore these hierarchal relationships in more detail. The tags are connected hierarchically via their spatial relationship – 'contained by' (Fig6). The number next to each tag name in Fig.6 show how confident the model is that the tag is not noise. The visualization is available as an applet at [23].

TABLE II. RESULT OF CLASSIFICATION FOR SELECTED AND UNSELECTED CASES AT $100M^2$ WITH 0.5 CUTOFF VALUES

Observed		Predicted					
		Selected Cases			Unselected Cases		
		Ori_Class			Ori_Class		
		0	1	% Correct	0	1	% Correct
Ori_Clas	0	1697	112	93.8	751	37	95.3
s	1	410	549	57.2	172	223	56.5
Overall Percentage				81.1			82.3

TABLE III. EVALUATION OF LOGISTIC MODEL (TABLE I) AGAINST MAUL CLASSIFICATION AT LOWER LEVELS OF DETAIL (PROBABILITY CUT OFF VALUE IS 0.5 – THE SAME AS IN TABLE II)

	Class	0	1	% Correct
Scale: 500m2	**0**	220	29	88.35
	1	99	216	68.57
Scale: 1000m2	**0**	39	11	78.00
	1	21	111	84.09
Scale: 2000m2	**0**	1	2	33.33
	1	0	47	100.00
Scale: 4000m2	**0**	0	0	
	1	0	10	100.00

IV. 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 Geospatial web has explored the use of flickr tags as a source for vernacular geography but there has been limited research in exploration of these images and tags at different levels of detail. This research has used data mining and text analysis techniques for selecting appropriate tags names as description of areas at different levels of detail. We have also presented a model used in post selection to calculate confidence (probability) values for each selected tag as a basis for assessing its likely correctness. The results were compared against manual inspection and it was observed that the range of the results correctly predicted by the model were from 80% at the most detailed level to 100% at the coarsest level. Future research will look into usage of clustering or road network partitions instead of arbitrary grid cells also threshold for selecting more than one tag for each region shall be addressed.

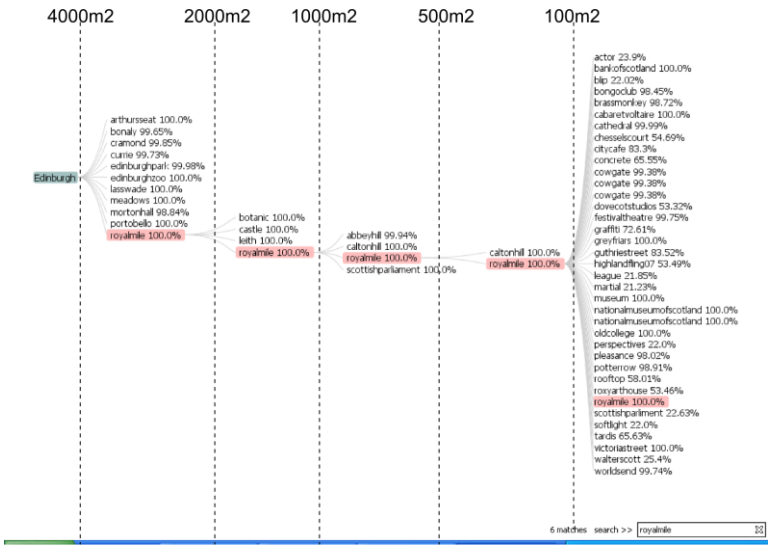

Figure 6. Tree view visualisation of selected tags and their confidence (probabilty) value as predicated by the model

V. REFERENCES

[1] S. Mustière and J. van Smaalen, "Database Requirements for Generalisation and Multiple Representations", in Generalisation of Geographic Information: Cartographic Modelling and Applications, W.A. Mackaness, A. Ruas, and L.T. Sarjakoski, Eds., Elsevier: Oxford, 2007. pp. 113-136.

[2] P. Agarwal, "Ontological considerations in GIScience". International Journal of Geographical Information Science, vol. 19, 2005, pp. 501-536.

[3] R.B. McMaster and K.S. Shea, "Generalization in Digital Cartography". Resource Publication in Geography: Washington D.C. 1992.

[4] R.S. Purves, et al., "The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet". International Journal of Geographical Information Science, vol. 21, 2007, pp. 717-745.

[5] C.B. Jones, R.S. Purves, P.D. Clough, and H. Joho, "Modelling vague places with knowledge from the Web". International Journal of Geographical Information Science, vol.22, 2008, pp. 1045-1065.

[6] P. Lüscher and R. Weibel. "Semantics Matters: Cognitively Plausible Delineation of City Centres from Point of Interest Data". in Proc. 13th workshop of the ICA commission on Generalisation and Multiple Representation. 2010. Zurich, Switzerland.

[7] L. Hollenstein and R.S. Purves, "Exploring place through user-generated content: Using Flickr tags to describe city cores". Journal of Spatial Information Science, vol.1, 2010, pp. 21-48.

[8] A. Scharl and K. Tochtermann, eds. "The geospatial web how geobrowsers, social software and the Web 2.0 are shaping the network society". Springer: London, 2007

[9] M.F. Goodchild, "Citizens as sensors: the world of volunteered geography". GeoJournal, vol. 69, 2007, pp. 211-221

[10] M. Zook, M. Graham, T. Shelton, and S. Gorman, "Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake". World Medical & Health Policy, vol. 2, 2010, pp.7-33

[11] R.S. Purves, "Methods, Examples and Pitfalls in the Exploitation of the Geospatial Web", in The Handbook of Emergent Technologies in Social Research, S.N. Hesse-Biber, Ed., Oxford University Press: Oxford, 2011, pp. 592 -624.

[12] K.S. McCurley. "Geospatial Mapping and Navigation of the Web". in Proc 10th international conference on World Wide Web, Hong Kong: ACM, 2001

[13] Flickr. Available from: http://www.flickr.com/map/, 2011, Last accessed: 21 July 2011

[14] T. Rattenbury and M. Naaman, "Methods for extracting place semantics from Flickr tags". ACM Trans. Web, vol. 3, 2009, pp. 1-30.

[15] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, "Generating summaries and visualization for large collections of geo-referenced photographs", in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. ACM: Santa Barbara, 2006

[16] F. Girardin, F. Calabrese, F.D. Fiore, C. Ratti, and J. Blat, "Digital Footprinting: Uncovering Tourists with User-Generated Cotent". Pervasive Computing. vol. 7, 2008, pp. 36-43.

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

[18] O.Z. Chaudhry and W.A. Mackaness. "Utilising Partonomic Information in the Creation of Hierarchical Geographies". in Proc 10th ICA Workshop on Generalisation and Multiple Representation. 2007. Moscow, Russia.

[19] R. Pasley, P. Clough, R.S. Purves, and F.A. Twaroch, "Mapping geographic coverage of the web", in Proc. of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems. 2008, ACM: Irvine, California.

[20] B. Croft, D. Metzler, and T. Strohman, "Search Engines: Information Retrieval in Practice", Addison-Wesley: Boston, 2009

[21] P.D. Allison, "Logistic Regression Using the SAS System: Theory and Application". Wiley Interscience:New York, 2001

[22] D.G. Kleinbaum and M. Klein, "Logistic Regression: A Self-Learning Text". 3rd ed., Springer: New York, 2010

[23] O.Z. Chaudhry. vailable from: http://www.omairchaudhry.net84.net/City_Viz/TreeView_Confidence.html. 2011 Last accessed: 20 Nov 2011

[24] Python http://www.flickr.com/services/api/ Last accessed: 20 Nov 2011

 ISBN: 978-1-61208-178-6