Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Semantic analysis of Citizen Sensing, Crowdsourcing and VGI

Alexis Comber
University of Leicester
Leicester, UK
ajc36@le.ac.uk

Sven Schade
JRC
Ispra, Italy
sven.schade@jrc.ec.europa.eu

Linda See
IIASA
Laxenburg, Austria
see@iiasa.ac.at

Peter Mooney
NIUM
Maynooth, Eire
peter.mooney@nuim.ie

Giles Foody
University of Nottingham
Nottingham, UK
giles.foody@nottingham.ac.uk

Abstract

This paper describes a semantic analysis of terms used to describe citizen sensing and crowdsourced data use in scientific analyses. It applies a latency analysis to journal abstracts downloaded from Scopus that matched one of number of terms related to crowd sourced data and citizen science. The latency analysis shows how the terms associated with crowdsourcing are related and how they have evolved over time.

1 Introduction

Whilst there is a long tradition of members of the public recording and sharing information about the world we live in, recent developments in digital technologies have driven an explosion of crowdsourced data collection and creation. Due to connected, location enabled digital devices – smartphones, cameras, tablets, notebooks etc – citizens are able to capture and almost share spatially referenced information about all kinds of processes (see for example [3] [6] [8]) via many different types of platforms – the web, social networks, server host sites (e.g. Flickr for photographs) – as well as targeted activities such as OpenStreetMap [9] and Geograph. Thus, it is now relatively simple for citizens to capture and share information about the world they live in, both actively (e.g. via OSM creation) or passively (e.g. via mining of twitter feeds).

The recent high level of scientific interest in crowdsourced data is high for 2 simple reasons. First, the very high data volumes that are potentially available to the scientist, and second, the low cost of such data. That is, at the core of much of the current scientific interest is the possibility that crowdsourced data may be able to replace data collected under the designed experiment that is where data are collected under a formal experimental design that includes sampling strategies, stratifications, etc. However, the critical issue using crowdsourced data in this way relates to the quality of the data. This not only relates to the reliability of observations and their labeling - whether they truly describe the phenomenon under consideration, but also to the spatial distribution of the observations, which depends on the locations of the individuals volunteering the information. Thus the controls over what is recorded and where is recorded that are frequently addressed by pre-specified experimental designs and the establishment of data capture protocols are lacking in crowdsourced data.

The focus of this paper is to consider how conceptualisations of crowdsourced data have evolved over time. It analyses the semantics of 'citizen science' activities as reported in the scientific literature for the period 1990 to 2013 in order to understand the changes in the way that the scientific community use, conceive apply such data.

2 Analysis

A text mining analysis of the semantics used in research describing the analysis, acquisition and qualities of crowdsourced geographic information was undertaken. The abstracts of 10,441scientific papers, published between 1990 and 2013, that contained any of the 24 the terms listed in Table 1 in their title, keywords or abstract were downloaded from Scopus (note: these terms were selected as initial set to investigate – future work will extend and refine these).

Table 1. Search terms used to extract scientific papers form Scopus

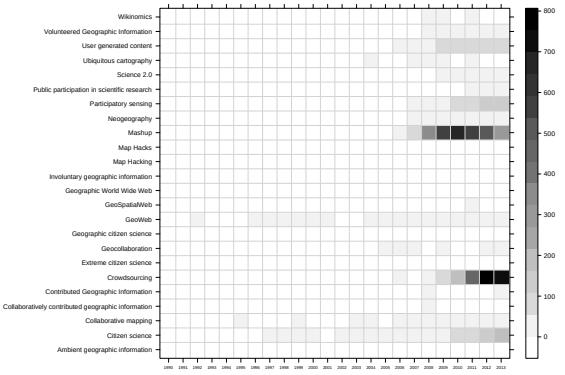
Terms
Science 2.0
Collaborative mapping
Wikinomics
Extreme citizen science
Geographic citizen science
Geocollaboration
Map Hacking or Map Hacks
Neogeography
Participatory sensing
Ubiquitous cartography
Mashup
Citizen science
Collaboratively contributed geographic information
Crowdsourcing
Geographic World Wide Web
GeoWeb or GeoSpatialWeb
Involuntary geographic information
Volunteered Geographic Information
Public participation in scientific research
Ambient geographic information
User-generated content
Contributed Geographic Information

A *Latent Dirichlet Allocation* (LDA), first proposed by [1] was used to analyse the content of the abstracts. LDA seeks to explain similarity in documents using unobserved, latent groups or *topics*. The idea is that each document includes a number of embedded topics which are indicated by the words that the documents contain and that the frequency of words in documents describe these associations. Latent approaches consider the data (documents) and the hidden concepts they contain (topics) from the standpoint of naivety and seek to determine the underlying similarities between documents and concepts. These techniques have been used in a number of spatial data analyses [11] [12] [4] [5] have applied them to integrate land cover data with different taxonomies. Here, citation data were downloaded from Scopus for publications that matched at least one of a number of search criteria.

The data were cleaned to remove English stopwords (conjunctions, pronouns etc.), numbers, punctuation, whitespaces and any words less than 3 characters long. The words were then *stemmed*. Stemming is the process of establishing common etymological roots for words such that, for example *propose* and *proposal* have the same stem of *propos*. The cleaned and stemmed abstracts were then organised into a *corpus* of 24 documents based on the year of publication.

The evolution of the terms and phrases related to citizen sensing listed above was analysed using the *term frequency-inverse document frequency* (tf.idf). The tf.idf weight is a commonly used in library sciences for document classification and information retrieval. It is a statistical measure and describes the importance of a word in relation to any given document. A frequency matrix was constructed describing the occurrence of each of the phrases in each of the 24 documents representing the corpus of abstracts for each year (1990 to 2013). This is shown in Figure 1 where the cells in the matrix indicate the number of times each term appears in each year. Note, that in this case corpuses were re-created for each year, no stemming or removal of stop words was performed, and search terms with more than one word were replaced with concatenated versions (e.g. such that "Citizen science" was replaced with "Citizen_science").

Figure 1: The frequency of occurrence for each search term.

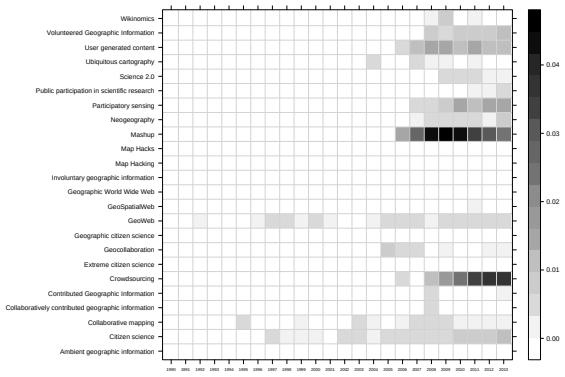


The terms in the matrix were weighted using the 'tf.idf' scheme described in [9]:

$$W_{ij} = \frac{n_i}{\sum n_i} ln \frac{D}{n_i}$$

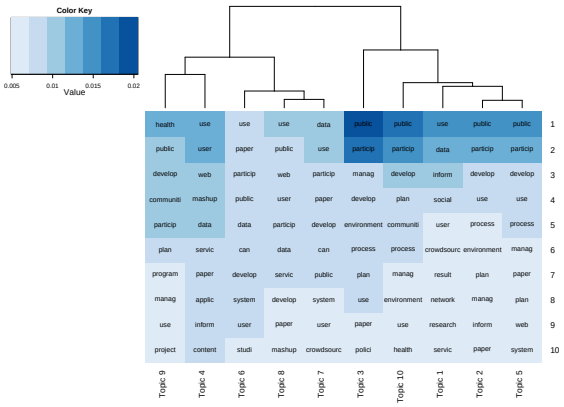
where W_{ij} is the weight of the i^{th} word in the j^{th} class, n_i is the number of times the word appears in the j^{th} class, Σn_i is the total length of the j^{th} class description, D is the total number of classes and n_j is the number of classes containing the i^{th} word. The weighting has the effect that a word that appears in all class descriptions has a zero weight, but a word appearing frequently in a few short classes has a high weight. The results of apply the are shown in Figure 2.

Figure 2. The changes in *tf.idf* values for the search terms 1990 to 2013.



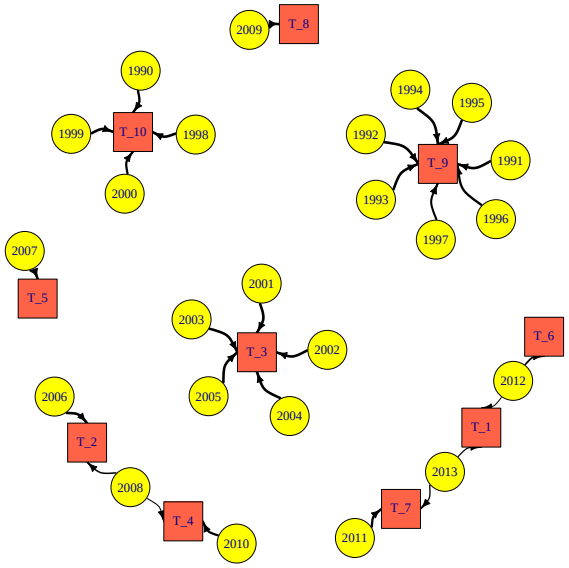
A Latent Dirichlet Allocation analysis was run on the corpus using the *topicmodel* package [2] in R, the opensource statistical software. Ten latent variables or topics were identified and these can be characterised by the terms that are most strongly associated with them from the posterior probabilities generated by the LDA of each term being associated with each topic (Figure 3). This suggests that there are 3 distinct topic groups: Topics 4 and 9 (*community*, *mashup*, *web*, *develop*, *health*) , Topics 6, 7 and 8 (*use*, *particip*, *web*, *public*) and Topics 1, 2, 3, 5 and 10 (*particip*, *develop*, *public*).

Figure 3. The 10 stemmed terms most strongly associated with each topic, shaded by the posterior probability of belonging to that topic and with the topics clustered.



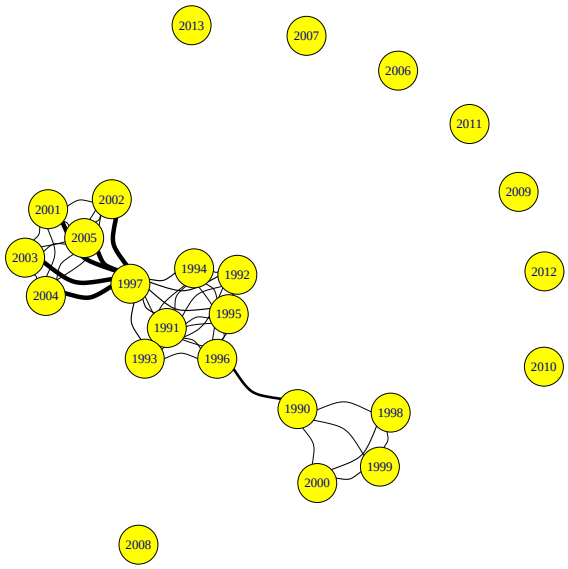
The LDA also generates posterior probabilities that each document is associated with each topic. These relationships between topics and documents via their semantics for each year can be visualised in a network, where the edges are defined by probability. For clarity the edges (connections) between years and topics (vertices) were removed if the posterior probability for each topic-year pair was less than the minimum posterior probability for that year plus the standard deviation [7]. The connections between topics and years is shown in Figure 4 and indicates an evolution over time of the concepts associated with publications in this domain.

Figure 4. The links between topics and years, with the strength of the link as defined by the posterior probability as determined by the LDA model indicated by the edge widths.



The connectedness between the semantics embedded in documents from different years is further illustrated in Figure 5. This shows the semantic distances between the documents for different years in the corpus. The recent explosion of publications, application and the wider discussion of the use of citizen sensed data in scientific publications are perhaps suggested by the lack of links between publications from more recent years compared to the 1990s and early 2000s – 1997 is particularly interesting year.

Figure 5. A network describing the semantic distances between documents published in different years.



3 Discussion Points

A number of areas for future consideration have been identified through this initial exploratory work. First, that the number of scientific papers that cite (not about) crowdsourcing topics has increased in recent years. Second that there are clearly identifiable evolutionary phases in the way that such information is referred to, witness the links in Figure 4 and Figure 5. These potentially reflects phases in GIS Science related to crowdsourcing between 1990 and 2005, the beginning of mashups, neogeography and so on in 2005-2006 seeing and a breadth of citizen science activities since then appearing to be disconnected. Thirdly, that recent research is clearly drawing from a much wider range of data sources, labelled in different and novel ways, potentially reflecting the rapid increase in the platforms and systems available to individual citizens that enable them capture and share a diverse range of different types of information, describing the world we live in. There are obvious areas for future research in considering who contributes such data, the impact of digital divides on the nature of the information that is contributed and potential biases towards western, developed populations and of course the nature of the technologies used to capture and share such information. On-going work is considering these issues

4 Acknoweldgements

This work was undertaken under the EU COST TD1202 ‘Mapping and the citizen sensor’.

References

[1] Blei D.M., Ng A.Y. and Jordan M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

[2] Grun, B and Hornik, K, 2013. Package 'topicmodels'. http://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf [available 07/01/14]

[3] Coleman, D., 2010. The potential and early limitations of volunteered geographic information. *Geomatica* 64 (2), 27–39.

[4] Comber, A.J., Fisher, P.F. and Wadsworth, R.A., (2007). Mining semantics of geographical information to generate user-relevant metadata, Spatial Data Usability Workshop, at *AGILE 2007, 10th AGILE conference on Geographic Information Science*, (eds. Monica Wachowicz and Lars Bodum), 8th May, Aalborg.

[5] Comber, A, Lear, A and Wadsworth, R, (2010). A comparison of text mining and semantic approaches for integrating national and local habitats data: semantic accuracy, error or inconstancy? In proceedings of Accuracy 2010. Pp297-300 in N Tate and P Fisher (eds.), *Proceedings of the 9th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, 20th -23rd July 2010, University of Leicester, Leicester.

[6] Haklay, M., Basiouka, S., Antoniou, V., Ather, A., 2010. How many volunteers does it take to map an area well? The validity of Linus' law to volunteered geographic information. Cartographic Journal 47, 315–322.

[7] Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013.

[8] Jones, C.E., Mount, N.J., Weber, P., 2012. The rise of the GIS volunteer. *Transactions in GIS* 16, 431–434.

[9] Mooney, P., Corcoran, P., 2012. The annotation process in OpenStreetMap. *Transactions in GIS* 16, 561–579

[10] Robertson, S.E. and Spärck Jones, K., 1976, Relevance weighting of search terms, *Journal of the American Society for Information Science*, 27(3), 129-46.

[11] Wadsworth, R.A, Comber, A.J. and Fisher, P.F., (2008). Probabilistic Latent Semantic Analysis as a potential method for integrating spatial data concepts, pp 99-108 in *Proceedings of the Colloquium for Andrew U. Frank's 60th Birthday*, (ed. Gerhard Navratil), GeoInfo Series 39, Vienna, ISBN 978-3-901716-41-6

[12] Wadsworth RA, Comber AJ, and Fisher PF, 2006 Expert knowledge and embedded knowledge: or why long rambling class descriptions are useful. In *Progress in Sptial Data Handling, Proceedings of SDH 2006*, (eds Andreas Riedl, Wolfgang Kainz, Gregory Elmes), Springer Berlin: 197 – 213