Assessing the Credibility of VGI Contributors Based on Metadata and Reverse Viewshed Analysis - An Experiment with Geotagged Flickr Images

Abstract

We present an approach to determine the credibility of content provided within a visual VGI source such as Flickr. We propose analysing the variability of selected user and photo metadata of geotagged Flickr photos with the location correctness of these images, which is our reference quality measure. These observed user and photo metadata can help to infer the credibility of contributors.

Keywords: Credibility, VGI, viewshed

1 Introduction

Volunteered Geographic Information (VGI) has shown an immense increase over the past decade. With massively increased production and availability of user generated geospatial data, considering the data credibility becomes a pressing issue. Flanagin & Metzger (2008) expressed the importance of assessing the subjective and objective nature of data credibility, which is a combination of *trust* and *expertise* (quality). Frew (2007) described how metadata about VGI can provide a basis for the judgement of quality of these data sources.

This work presents observations that help determining the user[1] credibility within visually generated VGI. These observations are derived based on an assessment of location correctness of visual VGI, which acts as a reference quality measurement for our study. In a series of steps, first the location of a described point of interest within a user-provided image is validated. This is done by testing whether that point of interest lies within the line of sight from where the image originates. We utilise geotagged Flickr images as an example, however this approach can be applied to other VGI sources as well. In a second step, we observe which photograph/user metadata can be utilised to infer the credibility of contributors regarding a correct geotagging. The approach is detailed in Section 2 and its analysis is described in Section 3.

2 Approach

We first implement a Flickr metadata crawler[2] that relies on the open Flickr API to fetch metadata of Flickr photographs for a specified set of tags. This enabled us to download metadata of photographs textually tagged with a particular point of interest, in case of this work, we use “Reichstag” and “Berlin”. The Reichstag is the German house of parliament, an attraction to many tourists travelling to Berlin.

To derive a reference quality measure for location correctness, a reverse viewshed for the point of interest is calculated. This determines the line of sight for our point of interest primarily based on surface elevation data. A *reverse* viewshed holds the same principles as a viewshed, however, it is utilised to determine the visibility of a given target point from many observer points (Fisher, 1996).

Subsequently, the geotagged photographs are overlaid with the reverse viewshed (Figure 1). Having this overlay in place, we are able to determine which photographs are textually tagged as "Reichstag” and “Berlin" *and* which are correctly geotagged within the range of visibility to the Reichstag. Photographs that are geotagged out of this visibility range are considered to either misrepresent the location from where the photograph was taken, or the photographed content represents something else other than the point of interest but tagged as the latter. Photographs belonging to either of these two groups are considered to be tagged with incorrect location. Figure 1 depicts some examples of incorrect geotagging and/or labelling. The example photo A in this figure is incorrectly geotagged since it lies outside of the line of sight, and incorrectly labelled since the object in the photo does not represent the Reichstag. Photo B is incorrectly geotagged but correctly labelled since it represents the Reichstag. Photo C is correctly geotagged but incorrectly labelled, as the image represents a sculpture from the soviet war memorial close by to the Reichstag. The visibility of the Reichstag from the positions of A, B, and C are further clarified using Google street view.[3]

Using this approach and the reverse viewshed analysis, we investigate which metadata of photographs (e.g., tag count of photographs) as well as metadata about users (e.g., the number of photos) correspond to the location correctness of photographs, acting as a starting point to eventually predict the credibility of photographers regarding correct geotagging. We achieve this through analysing the dependency relationship between those metadata and the location correctness of the geotags.

3 Results & Analysis

We manually study a sample of 182 geotagged Flickr images for the Reichstag in Berlin following the approach described in Section 2. Out of the 182 photographs, 25% (category **a**) are incorrectly geotagged as well as incorrectly labelled. 23% (category **b**) are incorrectly geotagged but correctly labelled. 22% (category **c**) are correctly geotagged but incorrectly labelled and 30% (category **d**) are correctly geotagged and labelled (Table 1). Figure 2 presents descriptive statistics of selected metadata elements for the four identified categories, which are the basis for the following analysis.

[1] Throughout this paper "user", "contributor" and "producer" refer to the same role

[2] Link to the source code: <ANONYMIZED>

[3] www.google.com/streetview

Regarding the average number of photos contributed by users to Flickr within each category reveal that producers of photos with incorrect labels have contributed significantly more photos over the years of their participation in Flickr (category **a** contributed 13,527 and c 11,585), as compared to categories **b** (3,887) and **d** (3,045).

Regarding the average tag count per photo, we see that producers within category **d** who have correctly geotagged and labelled the photos, have the least number of tags per photo (9), as compared to the total average of tags per photo within **a**, **b** and **c** (18). Both observation can be explained by bulk uploads done by users with high number of photos, which also explains the results of high tag counts to generalise their photo bulk.

Figure 1: Examples of Flickr images (green points) overlaid on the reverse viewshed. Red circle denotes the Reichstag

Table 1: Image classification concerning correct geotagging and labelling

Category	Correct Geotag	Correct Label
a (25%)	No	No
b (23%)	No	Yes
c (22%)	Yes	No
d (30%)	Yes	Yes

The average number of photo licenses reveal another interesting pattern. Photo licenses are optionally invoked by the photo contributors, to claim credit when others republish it, to protect from the creation of derivative work, or to (dis)allow commercial usage. Category **d** (0.7) have the highest average licenses per photo as compared to the total average of categories **a**, **b** and **c** (0.3). This also indicates a more careful dealing with images of category **d** users.

We also compute the distance to the target by taking the orthodrome between the geotag (as specified by the user) and the actual geographical coordinates of the Reichstag (as taken from Wikipedia[4]). This reveal that users within category **d** have on average the least distance to the target (300 m). Users within categories **a**, **b** and **c** have a distance to target varying between 700 and 900 meters. The closer to the point of interest a person is, the more focused the object would be in the image, thus, allowing the user to geotag/label more precisely. The further away from the point of interest, the user might become more imprecise when geotagging and labelling the image.

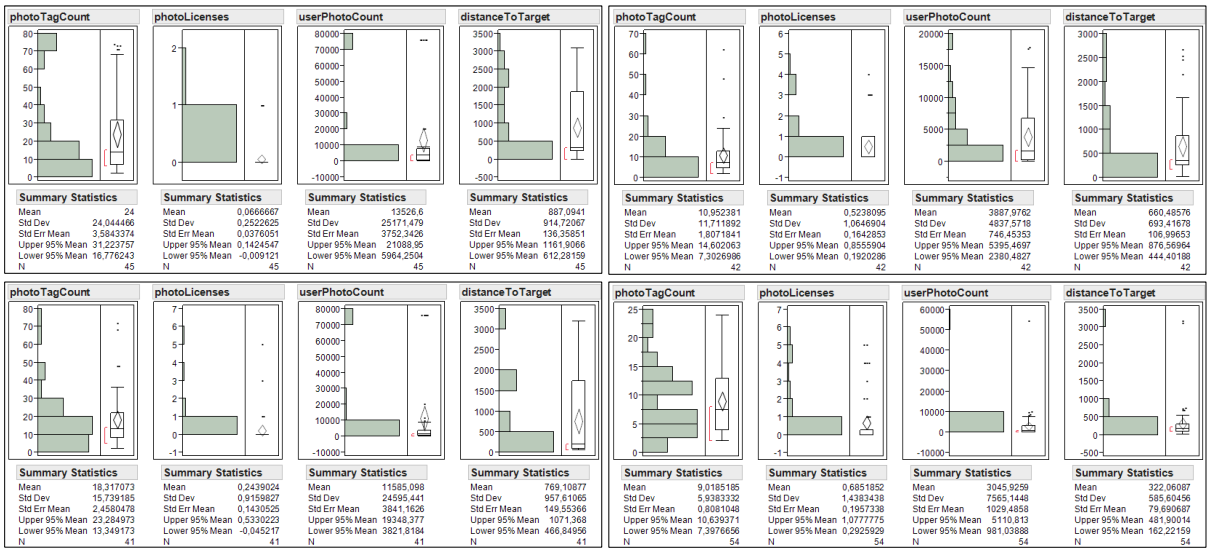
4 Conclusions and Outlook

We use a reverse viewshed to assess the location correctness of geotagged Flickr images for the Reichstag in Berlin. Based on this, we derive four categories of photos for (in)correct geotagging and labelling. Using the reverse viewshed as a reference quality measure, we analyse user and photo metadata on their variability within these four categories. We observe contributors who incorrectly label their photographs have on average the highest number of photos. Contributors who correctly geotag and label their photos have on average the highest number of licensed photos, the least distance to the target, and also the least number of tags per photo.

These observations are a starting point to heuristically assess expected image credibility relating to location and description correctness. In the future, we will refine our approach to a full prediction model. Considering content-based analysis functions and multivariate regression analysis could provide advanced quality predictions. We want to extend our studies to larger data sets and consider additional data sets from the VGI domain. These results will eventually enable new applications and improve drawing usage from mass VGI data.

[4] www.wikipedia.org

Figure 2: Distribution of data for each category. Left-right: a, b, c, d



References

[1]. Flanagin, A.J. and Metzger. M.J. 2008. The Credibility of Volunteered Geographic Information. GeoJournal, 72(3), 137-148.

[2]. Frew, J. 2007. Provenance and Volunteered Geographic Information. Online available: http://www.ncgia.ucsb.edu/projects/vgi/docs/position/Frew_paper.pdf (retrieved on 10.07.2012)

[3]. Fisher, P.F. 1996. Extending the Applicability of Viewsheds in Landscape Planning. Photogrammetric Engineering & Remote Sensing 62(11), 1297-1302.