Using Visual Analytics to Detect Problems in Datasets Collected From Photo-Sharing Services

Alexander Kachkaev*
giCentre, City University London

Jo Wood, *Member, IEEE*†
giCentre, City University London

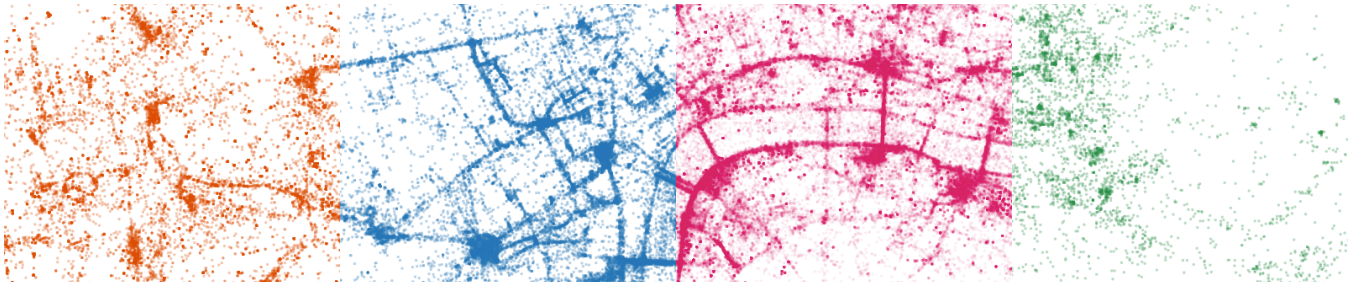

Figure 1: Distributions of geotagged photographs in Central London taken by users of three photo-sharing services between January 1st, 2008 and December 31st, 2011 and collected by means of APIs provided. *Left to right:* Picasa Web Albums, Panoramio, Flickr, Geograph.

Abstract

Datasets that are collected for research often contain millions of records and may carry hidden pitfalls that are hard to detect. This work demonstrates how visual analytics can be used for identifying problems in the spatial distribution of crawled photographic data in different datasets: Picasa Web Albums, Panoramio, Flickr and Geograph, chosen to be potential data sources for ongoing doctoral research.

This poster summary describes a number of problems found in the datasets using visual analytics and suggests that greater attention should be paid to assessing the quality of data gathered from user-generated photographic content.

This work is the first part of a three-year PhD project aimed at producing a pedestrian-routing system that can suggest attractive pathways extracted from user-generated photographic content.

1 Introduction

The quantity of user-generated content on photo-sharing websites and its availability has been given rise to various research studies for several years. The richness of the information that can be obtained from such services by means of their APIs encourage usage of photo-data as an input for a wide range of analysis, from examining contributors' behaviour [1] to measuring popularity and semantics of landmarks [2][4] or developing innovative trip planners for tourists [3][5][6]. The last group of projects is based on the ability of geotagged photo-content to measure place attractiveness and popularity and consider space as a set of spatially distributed points of interest.

Ongoing PhD research aims to construct a routing system based on user-generated photo content considering city space as continuos environment, and for this reason the accuracy, cleanness and representativeness of the input data is extremely important. Thus, more attention must be paid to potential sources and filtering, as there can be errors and bias in the contributors and spatial patterns of such data [7]. Picasa Web Albums[1], Flickr[2], Panoramio[3] and Geograph[4] were chosen as the candidate sources and were assessed for being suitable for the ongoing research by means of visual analytics.

The purpose of this work is to demonstrate the findings discovered when using visual analytics to examine the distributions of collected photo metadata.

2 Exposition

Interactive visual analytics software has been developed in order to explore collected sets and look for potential problems within them. Written using Processing and Java, it keeps metadata for hundreds of thousands entries in memory and allows panning, zooming, toggling layers, changing data representation, displaying statistics, etc. in real time. The tool helped to assess the datasets for the purpose of the research and detect a number of problems with the data.

Overall Density Evaluation The photo density view, where each item is represented by a semi-transparent circle and coloured by its source, is shown in Figure 1. In spite of the view's simplicity, it was found to be a useful instrument for assessing unexpected geographic patterns of photo density. For instance, with such visualisation it is clear that the street network is well-seen in case of Flickr and Panoramio, while barely apparent in Geograph, making it less attractive for this PhD research.

API Failure Detection Because service APIs are not necessarily designed for collecting vast numbers of photographs, they may fail in returning all of them for a requested geographic region. Such issues are normally fixed by splitting an area into smaller spatial units, but may not always succeed. Visualisations of the dataset itself and one of the API responses (Figure 2) helped to establish that Picasa caches results and returns photographs *outside* the requested bounding boxes in case if they are too small. This fact combined with an API limit of 1000 metadata entries per query made it impossible to get a representative distribution of photographs for the most popular places with Picasa API.

*e-mail: alexander.kachkaev.1@city.ac.uk
†e-mail: j.d.wood@city.ac.uk

[1] http://picasaweb.google.com/
[2] http://www.flickr.com/
[3] http://www.panoramio.com/
[4] http://www.geograph.org.uk/

Figure 2: Detection of a possible problem with Picasa API and its confirmation by means of visualising results of a single API request. Photos taken between January 1st, 2008 and December 31st, 2011 are orange, the rest (generally, most recent) are gray. *Right:* The server returned photographs outside the requested area.

Misplaced Photographs Detection The interfaces provided to users of photo-sharing services allow searching for a place and thus automatically put a photo on the map. According to the EXIF data attached to each photograph, at least 70% stored in Flickr have been geotagged manually rather than using GPS. Because the search queries can be inexact, some spots like the one in Figure 3 end up containing hundreds of photographs, not related to the local areas where they are placed. Such anomalies can negatively influence the results of a research and are easily discovered with visual analytics.

Figure 3: Anomaly containing unreasonably large numbers of photographs with the same coordinates. *Right:* Adding a randomly generated offset (standardised to fit a normal distribution) to photo coordinates and reducing alpha helps to find and see such places. *Background: © CC BY-SA OpenStreetMap and contributors.*

Comparison of Crawling Methods The visual analytics tool can be also used to compare various datasets, or even the same data collected differently. Figure 4 demonstrates that the choice of a crawling approach can significantly influence the distribution of the photographs.

Other Implications of the VA Approach Use of visual analytics to assess photo sets is not limited to the cases mentioned above. Other possibilities may include:

- Visualisation of distribution of photo illuminance based on EXIF data. This can help to see if a dataset contains significant numbers of photos taken indoors or overnight, which are not wanted for some analysis.
- Time filtering in order to observe dynamics in spatial-temporal distribution of the photographs. Such visualisation can help estimating the robustness of the dataset to events and seasonal changes.
- View of allocations of photographs with faces in them, with information obtained by means of service APIs or by doing image processing. This can give an idea of the amount of private photographs in the dataset, which are not useful in some cases.

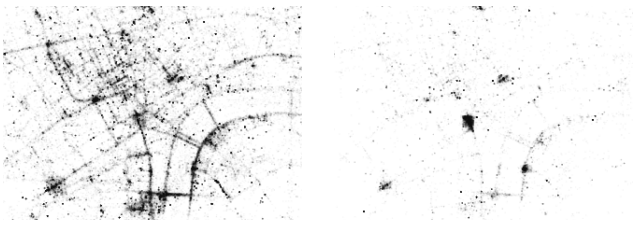

Figure 4: Photographs collected from Flickr (with the minimum edge of a bounding box ≈100m) using *left:* photographs only returned when using spatiotemporal requests and *right:* photographs only returned when using user-content requests (data kindly provided by Gennady Andrienko, Fraunhofer Institute, Germany).

3 Conclusions and Future Work

This paper demonstrates that a number of factors such as the nature of data, peculiarities of service APIs and even the interfaces users are dealing with when sharing their photographs can negatively affect the distribution of someone's input data and may lead to inaccurate research results. The work does not aim to provide a ready-to-use methodology to assess collected photographic information, but proposes paying greater attention to initial data analysis, problem detection and filtering before doing further analytical research. While the definition of a 'problem' in a photographic dataset will vary depending on the purpose of the research, the approach demonstrated here helps in understanding the effects of the data generation and collection process.

After potential concerns are detected, data filtering should take place. Visual analytics can be very useful during this stage too by helping to see the effectiveness of chosen filtering methods, thus playing a role of a feedback function. In future work in this project, the VA system will be developed to assess the effect of automatic filtering operations such as the removal of nighttime photography and photos of people.

References

[1] M. Clements, P. Serdyukov, A. P. de Vries, and M. J. Reinders. Using flickr geotags to predict user travel behaviour. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, page 851–852, 2010.

[2] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web*, WWW '09, page 761–770, New York, NY, USA, 2009. ACM.

[3] M. De Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu. Constructing travel itineraries from tagged geo-temporal breadcrumbs. In *Proceedings of the 19th international conference on World wide web*, page 1083–1084, 2010.

[4] S. Kisilevich, D. Keim, N. Andrienko, and G. Andrienko. Towards acquisition of semantics of places and events by multi-perspective analysis of geotagged photo collections. *GeoCart*, 2010.

[5] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura. Travel route recommendation using geotags in photo sharing sites. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, page 579–588, 2010.

[6] X. Lu, C. Wang, J.-M. Yang, Y. Pang, and L. Zhang. Photo2Trip: generating travel routes from geo-tagged photos for trip planning. page 143. ACM Press, 2010.

[7] Purves, R., Edwardes, A., and Wood, J. Describing place through user generated content. *First Monday*, 16(9), Sept. 2011.