
This is the unspecified version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/1320/

Link to published version:

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.
Using Visual Analytics to Detect Problems in Datasets Collected From Photo-Sharing Services

Alexander Kachkaev∗
igiCentre, City University London

Jo Wood, Member, IEEE†
igiCentre, City University London

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 to measuring popularity and semantics of landmarks or developing innovative trip planners for tourists. 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 continuous 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. Picasa Web Albums, Flickr, Panoramio and Geograph 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 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.
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.

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