Tile Based Visual Analytics for Twitter Big Data Exploratory Analysis

Daniel Cheng, Peter Schretlen, William Wright
Oculus Info Inc.
Toronto, Canada

e-mail: {dcheng, pschretlen, bwright} @oculusinfo.com

Abstract—New tools for raw data characterization of “big data” are required to suggest initial hypotheses for testing. The widespread use and adoption of web-based maps have provided a familiar set of interactions for exploring abstract large data spaces. Building on these techniques, we developed tile based visual analytics against a 1B point Twitter dataset and performed a Tukey-inspired exploratory data analysis.

Keywords-big data; visual analytics; exploratory data analysis

“Big data” refers to datasets that are so large that traditional approaches to process them are impractical or fail altogether. New tools for raw data characterization of these datasets through exploratory data analysis (EDA) [1] are required to suggest initial hypotheses for testing. The widespread use and adoption of web-based maps have provided a familiar set of human computer interactions for exploring abstract large data spaces [2]. Deriving from these techniques, we developed tile based visual analytics for Twitter datasets to perform John Tukey inspired EDA.

The Twitter dataset was composed of two collections. The first (Twitter1) was a curated dataset of 300 million geo-tagged records. The second (Twitter2) was a collection of one billion raw tweets without consistent geo-tagging. We use interactive plots of all the data to examine the entire dataset for initial hypotheses to test. Analysts can navigate from overviews to the lowest level of detail. As shown in Fig. 1, this EDA “living with the data” allows structures and emergent characteristics in the dataset to surface.

A tweet location heat map made anomalous visual artifacts immediately visible, including 1) data density in and around Japan was significantly higher than the rest of the world, suggesting that those data points were collected using a different method; 2) repeating vertical lines spaced at regular, ten degree intervals suggest a region for further investigation; 3) zooming in on Japan showed horizontal lines at sequential time indices, artifacts of tweets from traffic cam bots.

Plotting tweets between users as edges revealed numerous tweets originating from Indonesia that were sent to a location geo-located at (0,0), a region of ocean off the coast of Africa. As there are no manmade structures there, the anomaly suggests that there are errors in the data or collection method.

Tile-based visual analytics also affords richer insights than geo or cross plot heat maps. By combining a heat map with aggregation marker techniques, we are able to produce a visual analytic that allows the exploration of emerging tweet trends for dynamic geographic regions for instance.

![Figure 1. Anomalous visual artifacts: repeating vertical lines at 10 degree intervals (left). Tweets sent to geo-location 0.0 (right).](image1.png)

![Figure 2. “Tile apps” with each tile an independent visual analytic.](image2.png)

The tile analytics method divides the data into evenly sized hierarchical tiles, and then calculates and overlays a tile-bounded analytic, for example, the top five hashtags for the region. These analytics serve as “aggregated markers”. As the user zooms into the map, the hashtags shown become increasingly localized to their bounded region. Next steps include applying multiple analytics across multiple tiles. Fig. 2 shows examples.

In acknowledgement, this study was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract Number FA8750-12-C-0317. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of Defense position, policy, or decision.