Tile Based Visual Analytics for Twitter Big Data Exploratory Analysis

Daniel Cheng, Peter Schretlen, Nathan Kronenfeld, Neil Bozowsky, William Wright
Oculus Info Inc.
Toronto, Canada
e-mail: {dcheng, pschretlen, nkronenfeld, nbozowsky, bwright} @oculusinfo.com

***Abstract* — New tools for raw data exploration and characterization of "big data" sets are required to suggest initial hypotheses for testing. The widespread use and adoption of web-based geo maps have provided a familiar set of interactions for exploring extremely large geo data spaces and can be applied to similarly large abstract data spaces. Building on these techniques, a tile based visual analytics system (TBVA) was developed that demonstrates interactive visualization for a one billion point Twitter dataset. TBVA enables John Tukey-inspired exploratory data analysis to be performed on massive data sets of effectively unlimited size.**

Keywords - big data; visual analytics; exploratory data analysis.

I. Introduction

"Big data" refers to datasets that are so large that traditional approaches to process them are inefficient, impractical or fail altogether. New tools for raw data characterization of these datasets through exploratory data analysis (EDA) [7] are required to suggest initial hypotheses for testing. The widespread use and adoption of web-based geo maps have provided a familiar set of human computer interactions for exploring abstract large data spaces [5]. Deriving from these techniques, a tile based visual analytics system was developed and applied to Twitter datasets to perform John Tukey inspired EDA.

II. Living with the Data

An example Twitter dataset used is composed of two collections. The first (Twitter 1) is a curated dataset of 300 million geo-tagged records. The second (Twitter 2) is a collection of one billion raw tweets over 15 months without consistent geo-tagging. Interactive plots of all the data allows examining the entire dataset for initial hypotheses. Analysts can navigate from overviews to the lowest level of detail. This EDA "living with the data" allows structures and emergent characteristics to be observed (Fig. 1).

III. Tile-Based Exploratory Data Analysis

The tile-based visual analytics (TBVA) approach to EDA draws many parallels to interactions in browser-based tiled geographic maps such as Google Maps, Bing Maps, or Open Street Maps. These systems pre-render multi resolution image tiles stored in a power-of-two pyramid using a pre-defined grid, served on demand to a web client [4]. Our approach abstracts the geographic map to a general plotting surface. The tiles of our tiled plot are data agnostic: they may display not only geographic data, but also text, points of a scatter plot, a time series etc.

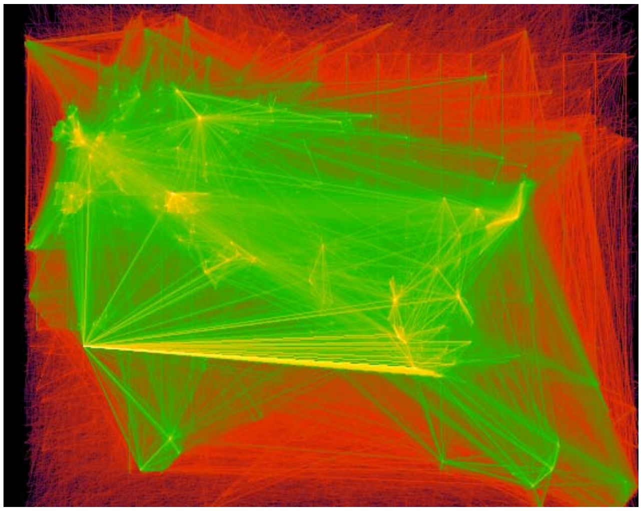


Figure 1. Anomalous visual artifacts: tweets at geo-location 0,0 (left), repeating vertical lines at 10 degree intervals (upper right).

TBVA tiled exploratory tools also differ from geographic tiled maps by deferring image rendering until request time. As shown in Fig. 2, TBVA generates tiles in three stages. The *aggregation* stage projects and bins data into a pre-defined grid, such as the Tiling Map Service [3] standard with a (z,x,y) coordinate system where z identifies the zoom level, and x,y identifies a specific tile on the plot for the given z. The *summarization* stage applies one or more summary statistics or other analytics to the data in each tile region storing the result in a data store. The *rendering* stage maps the summary to a visual representation, and renders it to an image tile at request time. This approach allows interactive modification and control of the visual representation. While any number of color ramps can be used to show density in this system, the preferred approach is to use a "spiral" ramp [8] that progresses through hues while increasing in luminance. The spiral ramp is perceptually efficient at presenting a sequence and for differentiating the greatest range of values.

The resulting pyramid of TBVA summary statistics allows interactive EDA through familiar map interactions like panning and continuous zooming, as well as interactive modification of the visualization, such as modifying color and shape mappings.

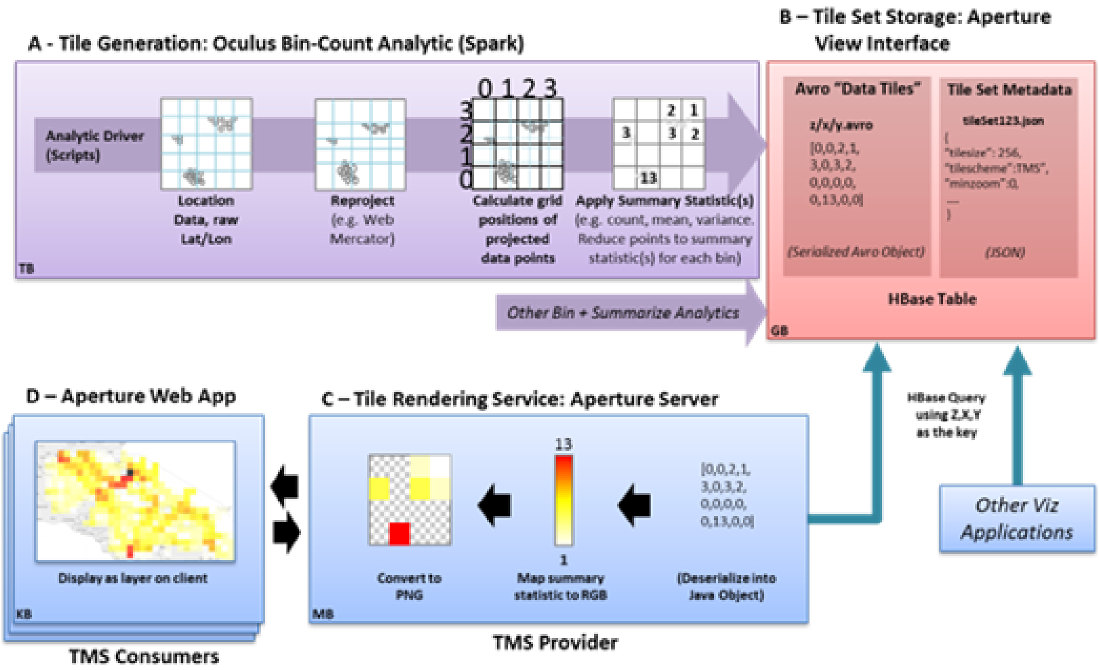

A - Tile Generation: Oculus Bin-Count Analytic (Spark)
Analytic Driver (Scripts)
Location Data, raw Lat/Lon
Reproject (e.g. Web Mercator)
0 1 2 3
3 2 1 0
Calculate grid positions of projected data points
2 1
3 3 2
13
Apply Summary Statistic(s) (e.g. count, mean, variance. Reduce points to summary statistic(s) for each bin)
TB
B – Tile Set Storage: Aperture View Interface
Avro "Data Tiles"
z/x/y.avro
[0,0,2,1, 3,0,3,2, 0,0,0,0, 0,13,0,0]
(Serialized Avro Object)
Tile Set Metadata
tileSet123.json
{ "tilesize": 256, "tilescheme":TMS", "minzoom":0, … }
(JSON)
HBase Table
GB
Other Bin + Summarize Analytics
HBase Query using Z,X,Y as the key
Other Viz Applications
D – Aperture Web App
Display as layer on client
KB
TMS Consumers
C – Tile Rendering Service: Aperture Server
Convert to PNG
13
1
Map summary statistic to RGB
[0,0,2,1, 3,0,3,2, 0,0,0,0, 0,13,0,0]
(Deserialize into Java Object)
MB
TMS Provider

Figure 2. TBVA tile generation and rendering workflow.

IV. Tiled Heatmaps

Using the TBVA approach, a tweet location density heatmap was created, with each tile representing a grid of 256x256 data bins. Each bin may contain a single value or a vector of data. This grid of bins is projected onto the plot, and a pyramid of count summaries is calculated for the region bounded by each bin.

The tweet location heat map made anomalous data artifacts immediately visible, notably that the 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. Zooming in on Japan revealed potential differences in data collection methods, as well as horizontal lines at sequential time indices. These are artifacts of bots on Twitter (Fig. 3).

The tweets have unique user ids, and plotting tweets in sequence for users as links or edges reveals hub and spoke travel patterns. In addition, numerous tweets originating from Indonesia were geo-tagged at latitude and longitude (0,0), a region of ocean off the coast of Africa, suggesting errors in the geo-tagging process or data collection method, as well as repeating vertical lines spaced at regular, ten degree intervals that suggest data quality issues (Fig. 1).

The same TBVA approach used to create the location density heatmap can also be applied to create a density scatterplot of any two dimensions in a dataset that has a large number of distinct values. Fig. 4 shows tiled scatterplots of two continuous numeric variables, sentiment and time. (Sentiment was calculated using an open source python library based on Wordnet [1].) A second example in Fig. 5 shows the frequency count of a categorical variable, hashtag, ordered by first occurrence, against time.

V. Tiled Density Strips

The tiled plotting approach is also useful for density strips, a form of one-dimensional heatmap [2]. Each vertical slice of a density strip corresponds to the value of a summary statistic. By plotting the values horizontally and applying a color ramp, the distribution of the summary statistic value for each dimension of the data is individually visualized. As with a tiled scatter plot, summaries are generated at multiple resolutions and stored in a power-of-two pyramid. Pan and zoom interactions are used to navigate the density strip providing a quick overview of the data as well as suggestions for points for more detailed, drill-down analysis. Fig. 6 illustrates interactive drill-down analysis with a density strip of the timestamp column of the Twitter 2 data collection.

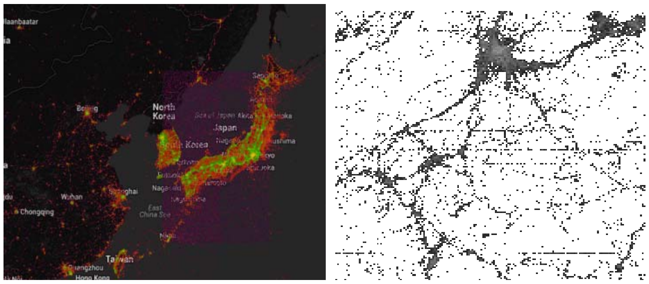


Figure 3. Heatmap over Japan: possible differences in data collection methods (left), horizontal lines created by Twitter bots (right).

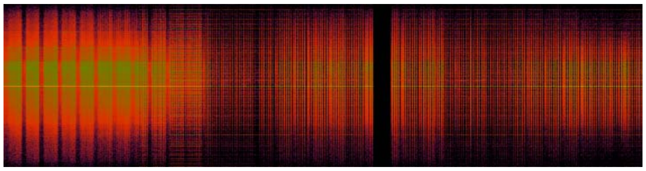


Figure 4. Tiled scatterplot example for the Twitter 1 collection showing sentiment vs. time.

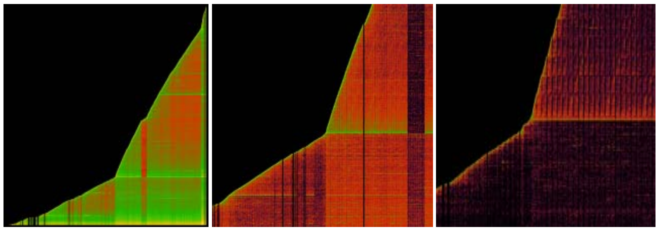
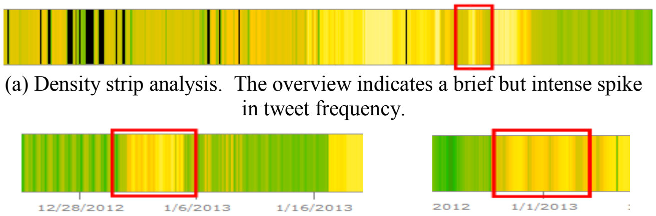


Figure 5. Hashtag occurences (vertical axis) over time (horizontal axis), with the hashtags ordered by time of first occurrence on the Y axis. Zooming in (left to right) to knee in curve reveals variations in data collection frequency and methodology.



(a) Density strip analysis. The overview indicates a brief but intense spike in tweet frequency.

b) Left. Zooming in reveals structure and local temporal distribution. Right. Zooming further shows tweet frequency at New Years 2013.

Figure 6. Drill-down analysis using a density strip.

VI. "Tile Apps" with Each Tile an Independent Visual Analytic

The TBVA also enables richer visual analytics than simple density strips and heat maps. Computational functions can be applied to each tile's data set. For example, a variety of aggregation marker "apps" can be used to represent large numbers of records in each tile region, promoting sense-making while keeping display complexity low [6]. For the Twitter data set, a heat map with tiled aggregation markers was combined, resulting in a visual analytic that allows the exploration of emerging tweet trends for dynamic geographic regions (e.g. most frequent, time series, change over two time periods). Like tiled plots, this approach divides a dataset into a tile pyramid, but instead of applying a statistical summary per bin, we calculate a tile-bounded analytic and overlay a visualization of the result. As the user zooms into the map, the analytic shown becomes increasingly localized to the bounded region.

Fig. 7 shows examples of Twitter hashtag analytics applied to tile regions and displayed as tiled aggregation markers.

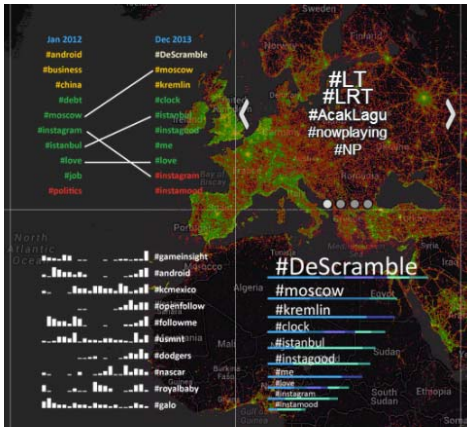


Figure 7. Examples of "Tile App" aggregation markers applied to hashtags occurences in the Twitter 1 collection. Analytics are applied to the tile region and a visual summary of the result is overlaid on the tile.

VII. Conclusion

Using a one billion point Twitter data set, it has been demonstrated that TBVA enables John Tukey-inspired exploratory data analysis to be performed on massive data sets of effectively unlimited size. Expected next steps include developing multiple analytics across multiple tiles, including providing interactions with TBVA analytics.

Acknowledgment

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.

References

[1] Computational Linguistics & Psycholinguistics Research Center, http://www.clips.ua.ac.be/pages/pattern-en#sentiment.

[2] Jackson, C., **Displaying Uncertainty with Shading**, The American Statistician, 62(4):340-347, 2008.

[3] Open Source Geospaital Foundation, **Tile Map Service Specification**, http://wiki.osgeo.org/wiki/Tile_Map_Service_Specification.

[4] Potmesil, M., **Maps Alive: Viewing Geospatial Information on the WWW**, Computer Networks and ISDN Systems 29.8, 1997.

[5] Schretlen, P., et al., **Interactive Data Exploration with 'Big Data Tukey Plots'**, Proc. IEEE VIS 2013, Accepted.

[6] Shneiderman, B., **Extreme Visualization: Squeezing a Billion Records into a Million Pixels**, Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. ACM, 2008.

[7] Tukey, J., **We Need Both Exploratory and Confirmatory**, The American Statistician, Vol. 34, No. 1, February 1980.

[8] Ware, C., **Information Visualization: Perception for Design**, Third Edition, Morgan Kaufman, 2013.