Google Analytics Spatial Data Visualization: Thinking Outside of the Box

Wanli Xing[1], Rui Guo[2], Ben Richardson[1], and Thomas Kochtanek[1]

[1] School of Information Science and Learning Technologies,
University of Missouri, Columbia, MO 65211, USA
{wxdg5,benjaminrichardson}@mail.missouri.edu,
KochtanekT@missouri.edu
[2] Department of Civil and Environmental Engineering,
University of South Florida, Tampa, Columbia, FL 33620, USA
rui@mail.usf.edu

Abstract. This paper showcases a methodology to assist website managers in determining the influence of their websites in regard to a particular location. This is achieved through enhancing Google Analytics by supplementing it with outside data sources. Though GIS software namely ArcGIS, the approach allows for more comprehendible geospatial analysis while also presenting maps overlays that are easier to grasp than the tools currently offered by Google Analytics. The Truman Presidential Library website serves as a case study to explore the potentiality of this approach.

Keywords: GIS, Google Analytics, Visualization, decision making, library.

1 Introduction

Google Analytics (GA) is the most commonly used web traffic analysis tool with over 80% of websites that track and monitor traffic data utilizing GA services. GA provides a variety of analytic tools to describe web traffic usage that are accessed through the GA dashboard. Despite its many benefits, GA is limited in its functionality especially with regard to geospatial data. The most frequently used methods to describe users' traffic data use statistics and graphs are derived from basic features in the GA dashboard. As a result, when it comes to spatial distribution associated with visitor traffic, researchers that examine the efficacy of GA geography dashboard (Figure 1) (which only provides the number of hits over a particular region (Country, State, City, etc.) are just beginning to determine how this information can assist web managers in facilitating decision-making. According to Turner (2010), broad measures of website usage such as virtual visits can be useful, but do not adequately portray the bigger picture of website statistics as these numbers provide little insight or ability to measure website performance on deeper levels of understanding. Consequently, a more sophisticated measure of the geospatial distribution of visits needs to be developed in order to assist decision makers in effectively evaluating their website.

S. Yamamoto (Ed.): HIMI 2014, Part I, LNCS 8521, pp. 120–127, 2014.

In an attempt to address the geospatial potential of GA, Clifton (2010) introduced the geomap overlay technique for adding other dimensions into the visitors' geographical distribution map to deliver information visually, and in turn, provide a more comprehensive picture of the performance of the website.

Until recently, few studies have applied these techniques or explored their potential for enhancing the visualization of geospatial data. This could be due to the fact that there has not been a mature framework developed that can extract or exhibit additional outside data to work in conjunction with GA. For instance, Clifton's map overlay method has inherent limitations in that the dimensions he suggested are limited to the data factors already contained in GA. For example, geospatial visit distribution data for the United States can be cross-referenced with data from the search engines visitors are using (data already collected by Google Analytics) and then displayed in the same graph. This technique does not add outside dimensions into the GA geographical graph which could have significant impact on website stakeholders' decision-making. For instance, academic websites may need to evaluate usage over different school districts (Xu et al, 2010) and commercial sites might want to consider a website's influence on different demographics (age, race, or income scale etc.) in different areas (Kumar et al, 2009). Therefore, Clifton's map overlay methodology is limited by not adding additional outside factors to the graph that could enhance the potential to provide a meaningful diagnosis of web traffic data.

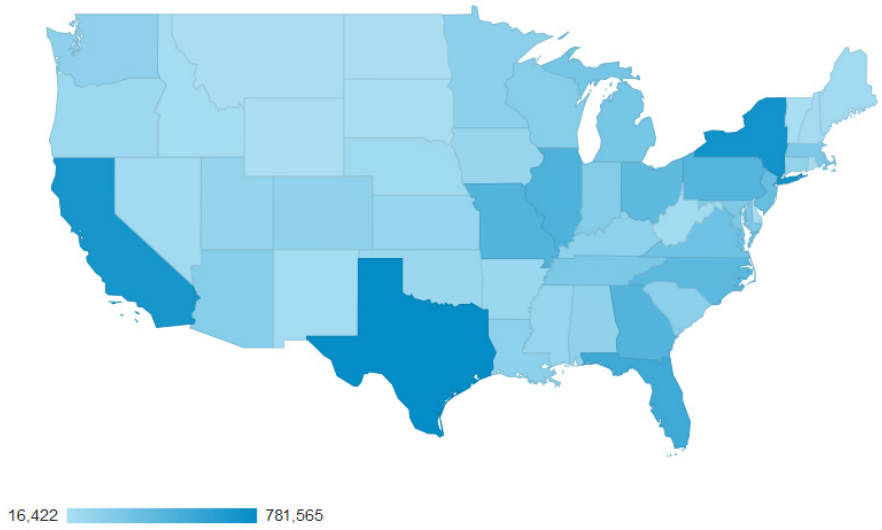

Fig. 1. Google Analytics nation-wide spatial distribution of Truman Library website visits dashboard

From an aesthetical and cognitive perspective, using a screenshot of the GA dashboard is insufficient because it merely presents a flat, color-coded graph. Various

researchers have indicated that enhanced aesthetics with different display formats can improve the perceivers' information processing dynamics from a design, psychological and practical perspective (Tractinsky, 2013 & 1997, Reber et al, 2004, Petersen et al, 2004). Moreover, the GA flat dashboard is incapable of displaying multi-measures. This makes the web managers' task of comprehending the graph more difficult as they attempt to process the overcrowded measures of information on the graph. In summary, current geospatial distribution in GA has three flaws: 1) no sophisticated metrics exist; 2) lack of empirical experience in adding other dimensions to the graph; 3) the graphical display is limited in the GA flat dashboard. Researchers using GA should explore new methods and tools for graphing measurements of visitors' spatial distribution and presentation.

To fill this gap, this study creates a new density metric to gain more insight into website usage and explores Geographic Information System (GIS) for visualization of web visitors' spatial distribution of a library website. This approach is capable of being employed with any website that utilizes GA.

2 Methodology

2.1 ArcGIS

In order to conduct spatial density analysis of a resource inventory, a Geographic Information System (GIS) needs to be implemented. GIS focuses on spatial data and the tools for managing, compiling, and analyzing that data. The distinguishing feature of a GIS is its capability of performing an integrated analysis of spatial and attributes data (Malczewski, 2004). This allows the spatial data to be manipulated and analyzed to obtain information useful for a particular application. Among many GIS software packages, ESRI's ArcGIS is one of the most popular types for working with maps and geographic information. The system provides an infrastructure for making maps and geographic information available throughout an organization, across a community, and openly on the Web. ArcGIS has a powerful functionality for map manipulation and aesthetics design by using ArcMap or ArcScene, whcih are ArcGIS Desktop application programs. The distinction between the two applications is that ArcMap is one of the core applications delivered with all licensing levels of ArcGIS Desktop, while ArcScene is part of the ArcGIS 3D Analyst extension.

Scope Selection

In order to perform analysis and visualization using ArcGIS, it is first necessary to initiate the raw data preparation for spatial analysis. Google Analytics allows users to export data to a file according to granularity. Depending on the needs, data can be exported out by country, continent, sub-continent, state, city and metropolitan etc. for further processing. Relying on the characteristic of the website, various study scopes can be defined. For instance, if the main target users of the website are located in the United States, then the United States can be chosen as the primary study scope to compare the distribution features between various states.

Supplementing Data

In undertaking any GIS-based work, the most common sources of spatial data should be collected to generate the base map. This is a significant step to introduce the outside source to Google Analytics. The source data includes data for layer generation and data for computation. In our study, we mainly supplement the TIGER/Line Shape-file data and population information for primary and secondary study scope, respectively.

Density Computation

In addition to the total visits by location directly obtained from Google Analytics, the visits density (representing the number of visits per person at one place), is computed for comparison. This outside population data is introduced to add an additional measure. While the common practice for web managers is to assess the influence of their companies or organization over a particular region this is merely based on the number of visitors from that particular place. To enhance this practice, we introduced the density measure to provide another perspective to measure such influence. Because it is natural that when a state has a larger population, it would have a much bigger possibility to have more visits. However, this alone does not necessarily mean that companies or organizations have more influence over that area. As an example, in terms of influence, 100 people heard of your organization in a city that has 1 million population does not necessarily outperform 50 people in a city which has 1000 population in total. Therefore, density for that place is calculated.

Assuming there are M representative years in the log datasets and website visits are cumulated through the total studied time period for different locations, density of website visits can be computed by the following equation.

$$Ds = \left(\sum_s \sum_{t=1}^{t=m} V_{ts}\right) / \left(\sum_s P_s\right)$$

Where, Ds denotes the density of website visits (visits/person) at location s, V_{ts} refers to website visits of the studied time period t at location s, and P_s is the population of location s. Here, s could be nation-wide, state-wide and city-wide etc.

Visualization

Using ArcGIS, multiple measures, density and visits to a particular location are displayed in the same graph.

2.2 Research Context

http://www.trumanlibrary.org

The Truman Presidential Library website as shown in Figure 2 was chosen for this research. The Truman Presidential Library website serves as a portal to online resources pertaining to the life of Harry Truman as well as providing information about visiting the Truman Museum. As such, it is a popular website that attracts many visitors from around the United States and worldwide that generates thousands of hits per day.

Fig. 2. Truman Library Website

2.3 Dataset

As an example, we experimented with this methodology to compare Truman Library's influence over different states. Since Google Analytics collects all the data for us automatically, we just simply downloaded the data by state. To compute density of visits, the 2012 population data for each state (except Alaska and Hawaii) of the United States was used. Additionally, TIGER/Line Shape-files by different layer types (e.g., block groups, census tracts and school districts) were collected from different online open sources. In order to facilitate the display of the graph and explanation, each state is assigned a number as table 1. Moreover, Figure 3 shows the visits for each state in each year.

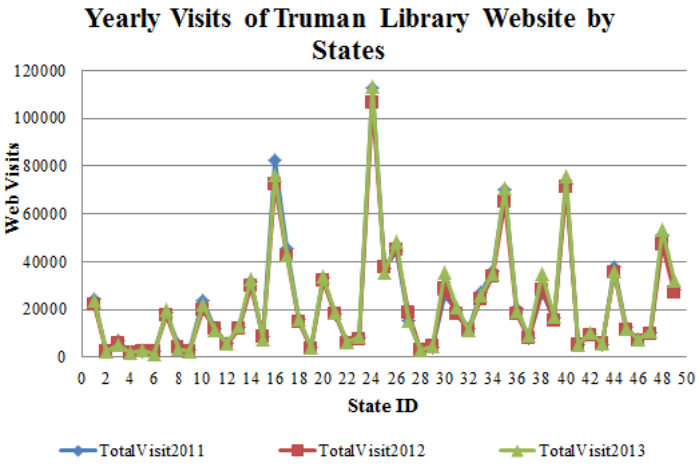

Fig. 3. State Visits in Truman Library

Table 1. State ID

StateID	1	2	3	4	5	6	7
State Name	Washington	Montana	Maine	North Dakota	South Dakota	Wyoming	Wisconsin
StateID	8	9	10	11	12	13	14
State Name	Idaho	Vermont	Minnesota	Oregon	New Hampshire	Iowa	Massachusetts
StateID	15	16	17	18	19	20	21
State Name	Nebraska	New York	Pennsylvania	Connecticut	Rhode Island	New Jersey	Indiana
StateID	22	23	24	25	26	27	28
State Name	Nevada	Utah	California	Ohio	Illinois	District of Columbia	Delaware
StateID	29	30	31	32	33	34	35
State Name	West Virginia	Marylan d	Colorado	Kentucky	Kansas	Virginia	Missouri
StateID	36	37	38	39	40	41	42
State Name	Arizona	Oklaho ma	North Carolina	Tennessee	Texas	New Mexico	Alabama
StateID	43	44	45	46	47	48	49
State Name	Mississippi	Georgia	South Carolina	Arkansas	Louisiana	Florida	Michigan

3 Results

3.1 Density

By introducing various data sources, we computed the density for each state as in Figure 4 according to the formula we proposed above. From the density graph we could

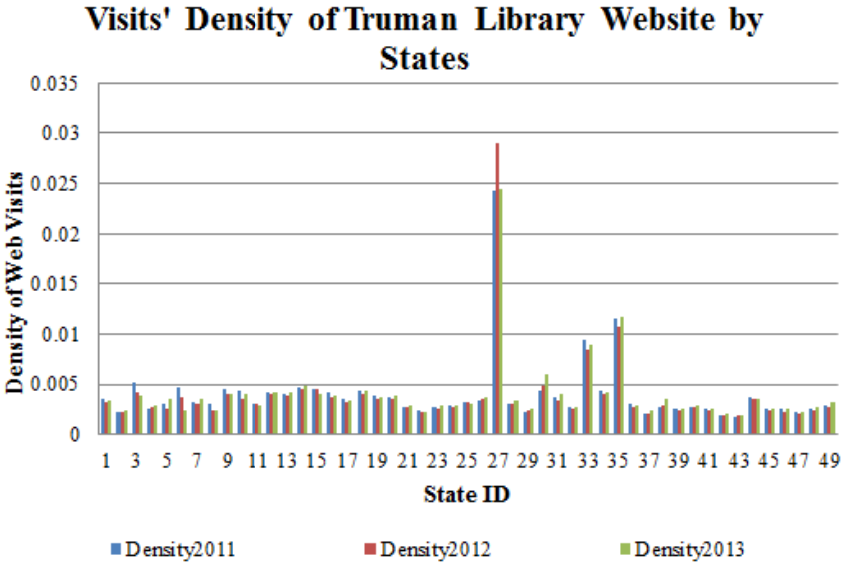

Fig. 4. Density Measurement

see which state density is the highest in combination with the State ID. In addition, in some regards, this density could also inform managers which state has the most influence.

3.2 Visualization

In order to assist mangers in their information processing, we implemented ArcGIS to visualize the information. As stated, ArcGIS is a powerful spatial data visualization tool and enables the visualization of a graph to hold multiple measures. In our context, we tried to present both the visits in each state and density measure in the same graph. As an example, we put the 2012 graph below.

As shown in the Figure 5, Texas and California had the most visitors, but in terms of density Missouri and DC have the highest value. This is understandable because Texas and California are the most populated states. However, Missouri is the location that hosts the Truman Library and DC is a location that supports many of the Truman Library activities. Therefore, Texas and California may have the most visits, but this does not necessarily mean Truman Library has the most influence over these places. In fact, it can be postulated from this data that the Truman Library has the most influence in Missouri and DC. In turn, density might be a better gauge to assess influence of the organization and company over a particular community.

In addition, compared with the Google Analytics flat graph, ArcGIS empowered maps can allow managers to process multiple-dimensional information simultaneously. While Google Analytics could only displey the visits to that particular location, ArcGIS is able to present various measures and information at the same time. Another significant aspect is to introduce the outside factor and measure to Google Analytics which is unattainable by using map overlay or similar techniques residing in Google Analytics.

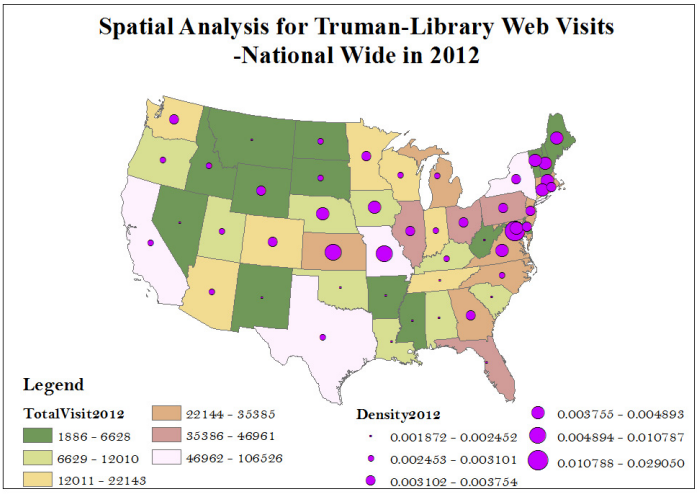

Fig. 5. ArcGIS Visualization

4 Conclusion

In sum, in this case study we utilized density metrics for measuring a website's influence over a region rather than merely depending on the numbers of visits to make conclusions. Further, we demonstrated that additional map layers and labels could be overlaid on an existing graph, and thus, present all of this information in one graph to provide a more refined analysis. It is hoped that website managers can utilize this approach to assist in their decision making process when examining the influence of their websites.

References

1. Clifton, B.: Reports Explained. In: Advanced Web Metrics with Google Analytics, pp. 97–127. Wiley. com (2012)
2. Kumar, C., Norris, J.B., Sun, Y.: Location and time do matter: A long tail study of website requests. Decision Support Systems 47(4), 500–507 (2009)
3. Malczewski, J.: GIS-based land-use suitability analysis: a critical overview. Progress in Planning 62(1), 3–65 (2004)
4. Petersen, M.G., Iversen, O.S., Krogh, P.G., Ludvigsen, M.: Aesthetic Interaction: a pragmatist's aesthetics of interactive systems. In: Proceedings of the 5th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, pp. 269–276. ACM (August 2004)
5. Reber, R., Schwarz, N., Winkielman, P.: Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? Personality and Social Psychology Review 8(4), 364–382 (2004)
6. Tractinsky, N.: Aesthetics and apparent usability: empirically assessing cultural and methodological issues. In: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, pp. 115–122. ACM (March 1997)
7. Tractinsky, N.: Visual Aesthetics. In: Soegaard, M., Dam, R.F. (eds.) The Encyclopedia of Human-Computer Interaction, 2nd edn. The Interaction Design Foundation, Aarhus (2013), http://www.interaction-design.org/encyclopedia/visual_aesthetics.html
8. Xu, B., Recker, M., Hsi, S.: The data deluge: Opportunities for research in educational digital libraries. In: Internet Issues: Blogging, the Digital Divide and Digital Libraries. Nova Science Pub Inc., New York (2010)