Visualizing User Behavior on the Places and Spaces Website

Avadhoot Agasti
School of Informatics and Computing, Bloomington, IN 47408, U.S.A.
aagasti@indiana.edu

Sharad Ghule
School of Informatics and Computing, Bloomington, IN 47408, U.S.A.
ssghule@iu.edu

Shreyas Rewagad
School of Informatics and Computing, Bloomington, IN 47408, U.S.A.
srewagad@iu.edu

Leonard Mwangi
School of Informatics and Computing, Bloomington, IN 47408, U.S.A.
lmwangi@indiana.edu

ABSTRACT

The *Places and Spaces: Mapping Science* exhibit introduces science mapping techniques to the general public and to experts across disciplines for educational, scientific, and practical purposes. The exhibit website *www.scimaps.org* provides information about people behind the exhibit; showcases maps and macroscopes; lists past, present and planned exhibit venues and dates. The website underwent a redesign in 2015 to update the organization and user interface of the website and this study aims to analyze and visualize the changes in behaviour of users of the website due to the redesign. After scraping raw data off of monthly website usage reports in HTML format and cleaning and trasnforming it into usable format, Tableau was used to create an interactive dashboard that would allow a user to gather insights from a number of visualizations. The dashboard allows a user to analyze the data from a high level as well as to drill down into details if required. Several interesting insights that were uncovered using the dashboard are presented.

KEYWORDS

Tableau, Visualization, Geospatial, URL, Hits, Pageviews, Parsing, Web Analytics, Line Graph, Tree Map, Dynamic Visualization

1 INTRODUCTION

Between the years 2005 and 2014, the *Places and Spaces: Mapping Science* exhibit worked towards the goal of bringing maps of science to the general public. In the year 2015, however, *Places and Spaces* made moves in a direction that marked both a continuation of and a development upon its past achievements. While the exhibit's first decade was mainly devoted to static maps of science, the second decade's mission is devoted to exploring the power and potential of macroscopes; which can be thought of as interactive tools to analyze complex, vast and slow phenomenons in the field of science. The website which acts as a source of information about the exhibit has a visitor base aross the globe and hosts a lot of informational content; videos, games and many science maps. The usage statistics of the website for the period of ten years from 2007 to 2017 are available through Webalyzer reports in HTML format. Visualizing all the information in these HTML files through an interactive dashboard would provide insights about the chnages in web traffic on the website after it was redesigned. Python was chosen as a tool to scrape the data from 120 HTML files and and seggregate it into a set of tables which would then be used to create visualizations. The visualizations include descriptive statistics of visitor demographics, page visits, content downloads; geospatial origin of visits; correlation between exhibit events and user activity.

2 CLIENT REQUIREMENTS AND VISUALIZATION GOALS

The intent of the client was to understand how the website is used in order to understand the audience and also to quantify the impact of the website. The requirements from the client helped us direct our analysis and visualizations towards answering below key questions:

- What is the geographic origin of the users?
- Are the majority of the users humans or crawlers and search engines?
- Is there a correlation between events and web-site traffic?
- Are users downloading content, if so, what are they downloading?
- What are users searching for?

Each of the visualizations we created provides insights that enable a user to answer a specific question. We attempted to make each visualization as interactive as possible and visually pleasing while conveying information in the best way possible.

3 TECHNICAL SOLUTION

The data pipeline for creating the visualizations has below important steps as explained in 1.

- Acquire the website usage data: The scimaps.org uses the Webalizer tool [3]. Webalizer analyzes the web server logs to create HTML report which provide various statistics of web site usage. While the actual web logs are available for only 2016 the Weblizer HTML reports are available for last 10 years(March 2017 to February 2017). We used these Weblizer HTML reports as source. The details of Weblizer reports, the exact HTML format and statistics captured is explained in the section 3.1.
- Combine the data in single data store which can be queried: Since the Weblizer HTML reports does not allow us to query the data, we required to convert them into a structured format. We decided to convert the HTML reports into comma-separated format (CSV). The section 3.2 explains the implementation of the parser program which converts data into CSV format. While designing the CSV format, we added metadata fields like year and month so that we can filter the data for a specific duration.

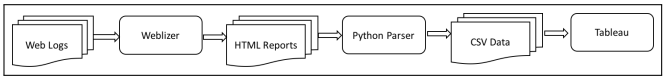


Figure 1: Data Pipeline.

Top 100 of 30517 Total URLs

#	Hits		KBytes		URL
1	10401	2.46%	170838	0.34%	/
2	3686	0.87%	364	0.00%	/robots.txt
3	3481	0.82%	100542	0.20%	/styles/css/PS_Global.css
4	1938	0.46%	359408	0.71%	/exhibit/docs/05-boyack.pdf
5	1787	0.42%	3078	0.01%	/home/panel
6	1536	0.36%	139776	0.28%	/Scripts/jquery-1.11.1.min.js
7	1504	0.36%	73984	0.15%	/Scripts/jquery.cycle.all.js
8	1489	0.35%	2776	0.01%	/Scripts/SlideshowBanner.js
9	1338	0.32%	51183	0.10%	/scimaps/atlas_of_science.html
10	1183	0.28%	15151	0.03%	/contact/
11	1070	0.25%	32604	0.06%	/iteration
12	871	0.21%	19405	0.04%	/advisory_board.html
13	854	0.20%	13356	0.03%	/what_is_a_science_map.html
14	814	0.19%	1154722	2.29%	/docs/EXHIBIT_MASTER_BOOKLET.pdf
15	651	0.15%	160	0.00%	/css/zoommap.css
16	624	0.15%	10025	0.02%	/home.html
17	536	0.13%	16570	0.03%	/browse_maps.html
18	526	0.12%	249328	0.49%	/exhibit/docs/Garfield1964use.pdf
19	512	0.12%	68722	0.14%	/exhibitions.html
20	453	0.11%	7606	0.02%	/team.html
21	428	0.10%	14888	0.03%	/mapstore
22	422	0.10%	6729	0.01%	/iteration/10
23	421	0.10%	15002	0.03%	/ambassadors.html
24	411	0.10%	1427781	2.83%	/docs/Kids_map_key.pdf
25	411	0.10%	632946	1.25%	/docs/PS_AnnualReport_2013_web.pdf

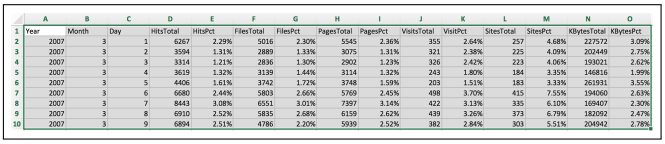
Figure 2: Sample Webanalyzer Report.

- Upload the data in visualization tool: We used Tableau [2] as our visualization tool for the visualizations. Tableau supports importing CSV data.
- Create individual visualizations: We created multiple reports in Tableau to satisfy various project requirements. Each Tableau report tries to answer a group of requirements for the project. Each report follows a similar pattern of filtering the data so as to maintain consistency across all reports. 4 section explains each visualization in detail.
- Create a single dashboard by combining all visualizations: While it is useful to analyze each dataset separately, it also helps to get a combined view of the overall website usage. We created dashboard from all the visualizations which helps in analyzing all the website usage data in one go. The dashboard provides interactive filters using which user can slice and dice data and analyze the usage pattern effectively.

3.1 Source Data

As explained in section 3, the Webanalyzer reports in HTML format are used as source data. These reports are available for last 10 years on monthly basis. Each report has following sub-sections

- Monthly statistics



Year	Month	Day	HitsTotal	HitsPct	FilesTotal	FilesPct	PagesTotal	PagesPct	VisitsTotal	VisitPct	SitesTotal	SitesPct	KBytesTotal	KBytesPct
2007	3	1	6267	2.29%	5016	2.30%	5545	2.36%	355	2.64%	257	4.68%	227572	3.09%
2007	3	2	3594	1.31%	2889	1.33%	3075	1.31%	321	2.38%	225	4.09%	202449	2.75%
2007	3	3	3314	1.21%	2836	1.30%	2902	1.23%	326	2.42%	223	4.06%	193021	2.62%
2007	3	4	3619	1.32%	3139	1.44%	3114	1.32%	243	1.80%	184	3.35%	146816	1.99%
2007	3	5	4406	1.61%	3742	1.72%	3748	1.59%	203	1.51%	183	3.33%	261931	3.55%
2007	3	6	6680	2.44%	5803	2.66%	5769	2.45%	498	3.70%	415	7.55%	194060	2.63%
2007	3	7	8443	3.08%	6551	3.01%	7397	3.14%	422	3.13%	335	6.10%	169407	2.30%
2007	3	8	6910	2.52%	5835	2.68%	6159	2.62%	439	3.26%	373	6.79%	182092	2.47%
2007	3	9	6894	2.51%	4786	2.20%	5939	2.52%	382	2.84%	303	5.51%	204942	2.78%

Figure 3: Daily Statistics Sample Records.

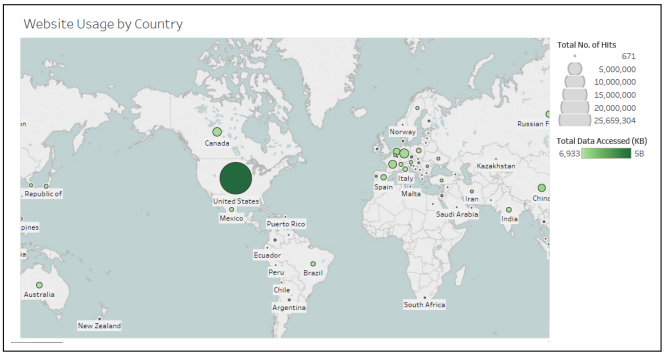


Figure 4: Geospatial Analysis of Web Traffic

- Daily statistics
- Hourly statistics
- Top 100 URLs
- Top 10 entry pages
- Top 10 exit pages
- Top 30 referring Sites
- Top 20 search strings
- Top 15 user agents
- Top 10 countries

Each section in webanalyzer report has HTML table. Figure 2 explains sample table from webanalyzer HTML report.

3.2 Data Parser

As explained in section 3, each Webanayzer HTML reports is converted into CSV format. We implemented Data Parser Python script which scrapes the Webanalyzer HTML report and converts it into CSV structure. The data parser uses Python module called BeautifulSoup to parse the HTML. It then iterates over all 'A' tags to find the section header within HTML report. Finally it iterates over the HTML table elements consisting TR and TD tags to extract the data and writes it in CSV file. The data parser code is available at [1] repository. The figure 3 shows sample records from the DAYSTATS.csv which is one of the output CSV created by the data parser.

4 VISUALIZATIONS

In this section, we explain various visualizations we created to satisfy the project requirements. Section 4.7 provides one page interactive view of overall statistics whereas the subsequent visualizations support detailed analysis of individual statistics.

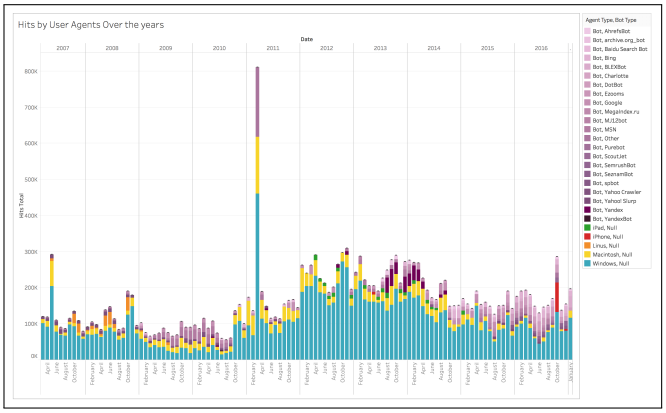


Figure 5: Accessing scimaps.org Traffic (2007 - 2017)

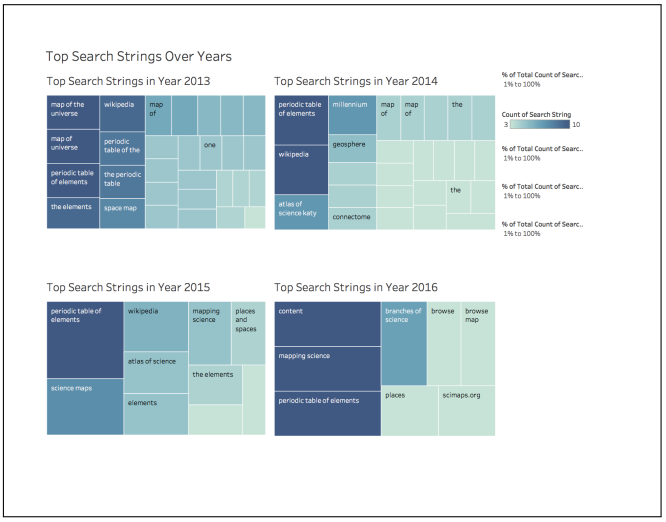


Figure 6: Top Search Strings in year 2013 to year 2016.

4.1 Top Countries and Trend

Using Tableau, we were able to geocode country names into lattitude and longitude values. We the visualized the same using one of the in-built maps in Tableau. We chose to use the proportional symbol map for this purpose that where the size of the symbol would indicate the number of total hits over a period of timme. Using color we were able to show levels of total data consumed by each of the countries. Users are predominantly from United States but one key insight that was uncovered is that after 2015, proportion of users from countries(e.g. Germany and France) apart from the US has increased.

4.2 Top Agents and Trend

Website traffic is mainly attributed to the various requests that are made to the server hosting the content. Let us review the requests made for scimaps.org. We have data of top 15 user agents from March-2007 to Jan-2017. Judging by the yearly pattern of the number of hits, we can mark 2011 as the pinnacle, after successive unremarkable hit fluctuations from 2007-10. there was a record 115% increase in the hits as compared to 2010. And the year 2012 was even better with a 21% increase over 2011. But since then the website has observed a downfall. The year 2013-14-15 have all marked the negative trend in the number of visitors/requests made for the site. Justifying a need for a website overhaul in 2015. We observe an increase of 10% in the request in the subsequent year. And the positive trend seems to be continuing as January records an increase of 26% over Dec-2016. Let us dive deeper and inspect the terminals that are making these requests. We have observed 6 types of terminals that attribute to the traffic on scimaps.org, the are:

- Windows PC
- Macintosh
- Linux Terminals
- iPhone
- iPad
- Bots

We see that the majority of traffic seems to be coming from Windows PC during the majority of 2007 - 2015. But, the hits/requests made by other terminals/bots seems to be catching up and the share evens out after the websiteâĂŹs overhaul. Till 2015 we observe that majority of traffic is from windows users, the traffic made by bots appears to have risen during this time. After the overhaul, Windows and bots share an approximate equal share of the website traffic, with mac traffic share increasing to 10.32% in 2017 over 2016. Of all the bots, major traffic should be attributed to Google, Yahoo Slurp and MSN bots from 2007 - 2010. And, in recent past DotBot and Bing seems to be the prime bots making 11.59% and 11% of traffic in 2016 respectively.

4.3 Top Referrals

<TODO: Leonard>

4.4 Top Searches and Trend

We used tree map visualization to plot the top search strings for every individual years. The tree maps for individual years, specifically 3 years before the website was reorganized and 1 year after the website was reorganized are placed side by side. This helps in understanding the trend of the search strings. Figure 6 provides the screenshot of the visualization. Please refer to the Tableau live implementation to see the interactive version of this visualization which showcases many details on the mouse-over.

The analysis of this visualization clearly identifies the trend. Before 2016, the maximum search strings were related to 'periodic table of elements' while in 2016 the focus is shifted towards 'mapping science'. However, these two topics are consistently amongst the top 10 searches throughout the analysis period.

4.5 Most Popular Pages

In this section we intend to discover the kind of content that is being consumed by the users on the scimaps.org website. The website observed a great boost in content consumed during 2008, 2011 and 2013 there was over 40%, 60% and 45% more content consumed as compared to the last year respectively. After the website revamp

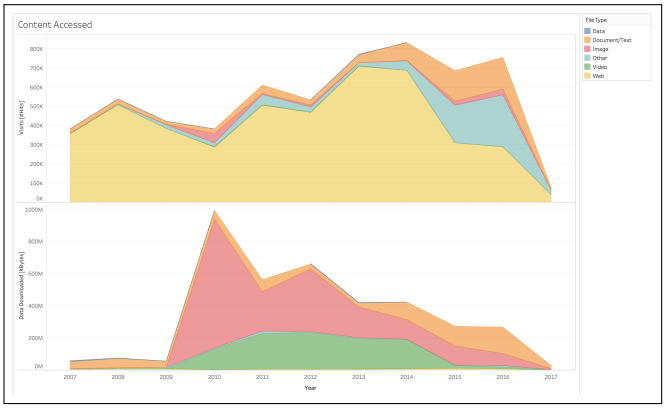


Figure 7: Distribution of Content Accessed (2007 - 2017)

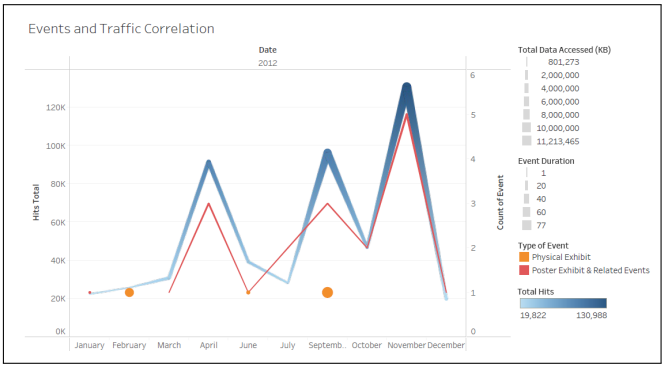


Figure 8: Correlation between Exhibit Events and Web Traffic

in 2015, we observe a rise in content consumption but the increase in not significant as the one observed during the above mentioned years. In terms of the downloaded content, 2010 observed a staggering 1678% increase in the data downloaded over 2009. But, the same level has not been maintained since then. Perhaps there might be some event that occurred during/before oct-2010 as we observe an increase of 43096% over Nov-2010. This should be further clarified from the events table. To attain an intuitive understanding of this we have broadly classified the website content into 6 categories:

- Data
- Document/Text
- Images
- Video
- Web
- Other

Following some take aways from our analysis of the user behavior on scimaps.org:

- Over the years there has always been a decrease in the amount of data consumed/downloaded.
- There has been a steady rise in the amount of documents browsed/downloaded over the years. The monotonic relationship doesnâĂŹt seem to have any effect of the website redesign as we observe a minor rise. A drastic change could have implied the userâĂŹs interest.
- Looking at the user behavior for images on scimaps.org, we see a clear interest. The images/visualizations showcased in the year 2010 marked the peak. There was over 3200% increase in the images viewed and 670,000% increase in the image data downloaded. This feat has not been repeated since then but we can see that the users are generally enticed by the image content that is being posted on the website. Since the website revamp in 2015, there has been a great increase user interest for images at scimaps.org.
- Taking a look at the video/media consumed by the users, we see a substantial increase in the videos downloaded and visited during 2010. This was the same case with images. This establishes a fact that the content that the website hosted in 2010 was highly appreciated. Since then we see a negative trend in the video content consumption and the website redesign doesnâĂŹt seem to have addressed that issue.
- The website redesign doesnâĂŹt seem to have positively affected the users web based content consumption.
- Looking at some of the miscellaneous content that the users been to consume at scimaps.org, we again see a huge rise in 2010. But, since then the downloads and visits seem to have declined. There has been a steady decline since 2012 till date, which indicates that the redesign seems to have no impact.
- Looking at the distribution of user visit and download pattern, we observe a trend. Earlier during 2007 -14 we observe huge web content consumed which drastically changed after 2015. This has been substituted by image, document and miscellaneous content.

4.6 Events and Web Site Traffic Corelation

In order to answer one of the key questions, we intended to visualize the trends in web traffic when there were exhibit events. We were able to get information about events in past ten years and classified the events as physical exhibits and poster events. The two datasets used for this visualization were combined using inner join in Tableau. There is clearly a correlation between events and web traffic as well as total data acccessed by the users. Physical events tend to affect web traffic by a significant amount and duration of the events plays a role as well.

4.7 Dashboard

<TODO: Leonard>

5 KEY INSIGHTS

TODO: Leonard

6 OTHER SIMILAR TECHNOLOGIES

There are various performance parameters and aspects that arenâĂŹt reviewed in the current paper due to unavailability of data. A good web analysis done by mining the user data would yields insights that could boost website traffic and potential business advancements. In this internet age the delivery of such Key Performance

Indicators and solutions need to be real-time. In this section we review some of the tools that could help in web analysis.

6.1 Google Analytics

This is a freeware made by Google to monitor and report website traffic. The tool showcases the descriptive statistics of the website on high-level. We can also procure intricate details and visualize the trend in user behavior pattern across the webpages. Apart from the fine grained details, the tool possesses Google intelligence and GoogleâĂŹs proprietary machine learning library integration. Thus making the Website owner aware of the possible actions that could potentially increase in website traffic. In order to link the tool to the website a tracking code is added to the web pages, this essentially integrates Google analytic in your website. This is also bundled along with other Google services which would help monitor the website via mobile devices.

6.2 Piwik

Piwik is a free open-source utility tool that can help the owner analyze the website traffic and determine the performance of various content hosted. Over the years users have added myriad plugins to Piwik to perform in detailed analysis and user profiling of the website traffic. This coupled with mobile integration helps the owner monitor the website traffic.

7 CONCLUSIONS

TODO: Leonard

A WORK DISTRIBUTION

The co-authors of this report worked together on the design of technical solutions, visualizations, implementation and documentation. Specifially, below given is the work distribution

- Avadhoot Agasti
  - Team lead and overall coordination.
  - Data parser implementation.
  - Visualization of top searches.
  - Putting together latext template for report writing.
  - Writing section 3 and section 4.4 in this report.
- Sharad Ghule
  - Visualization of top countries
  - Visualization of events and website traffic coorelation
  - Helping with intermediate deliverables
  - Writing Abstract, section 1, section 2, section 4.6 and section 4.1 in this report.
- Shreyas Rewagad
  - Visualization of top agents
  - Visualization of most popular URLs
  - Research on other similar technologies
  - Writing section 4.2, section **??** and section 6 in this report.
- Leonard Mwangi
  - Visualization of top referrals
  - Visualization of overall statistics
  - Creating dashboard
  - Writing section 4.3, section 4.7 and section 5 in this report.

B ACKNOWLEDGEMENTS

The authors thank Prof. Katy Borner for her technical guidance. The authors would also like to thank TAs of Information Visualization class for their valued support.

REFERENCES

[1] Avadhoot Agasti. 2017. ivmooc-scimap-webanalytics. (2017). https://github.com/avadhoot-agasti/ivmooc-scimap-webanalytics/ Online; accessed Apr-03-2017.
[2] TABLEAU. Tableau Desktop. (????). https://www.tableau.com/products/desktop Online; accessed Apr-03-2017.
[3] Weblizer. 2017. The Wenlizer: What is it. (2017). http://www.webalizer.org/ Online; accessed Apr-03-2017.