# 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

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

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

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

## ABSTRACT
TODO: Sharad

## KEYWORDS
ACM proceedings, LaTeX, text tagging

## 1 INTRODUCTION
TODO: Sharad

Lorem Ipsum is simply dummy text of the printing and type-setting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum.

## 2 CLIENT REQUIREMENTS AND VISUALIZATION GOALS
TODO: Sharad

Contrary to popular belief, Lorem Ipsum is not simply random text. It has roots in a piece of classical Latin literature from 45 BC, making it over 2000 years old. Richard McClintock, a Latin professor at Hampden-Sydney College in Virginia, looked up one of the more obscure Latin words, consectetur, from a Lorem Ipsum passage, and going through the cites of the word in classical literature, discovered the undoubtable source. Lorem Ipsum comes from sections 1.10.32 and 1.10.33 of "de Finibus Bonorum et Malorum" (The Extremes of Good and Evil) by Cicero, written in 45 BC. This book is a treatise on the theory of ethics, very popular during the Renaissance. The first line of Lorem Ipsum, "Lorem ipsum dolor sit amet..", comes from a line in section 1.10.32.

The standard chunk of Lorem Ipsum used since the 1500s is reproduced below for those interested. Sections 1.10.32 and 1.10.33

from "de Finibus Bonorum et Malorum" by Cicero are also reproduced in their exact original form, accompanied by English versions from the 1914 translation by H. Rackham.

## 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 [WeblizerWeblizer2017]. 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.
- Upload the data in visualization tool: We used Tableau [TABLEAUTABLEAU] 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.
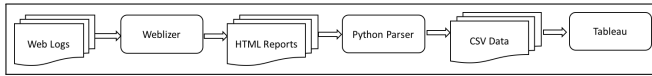
Figure 1: Data Pipeline.



Figure 2: Sample Webanalyzer Report.



Figure 3: Daily Statistics Sample Records.

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 [AgastiAgasti2017] repository. The figure 3 shows sample records from the DAYSTATS.csv which is one of the output CSV created by the data parser.

## 4 VISUALIZATION

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.
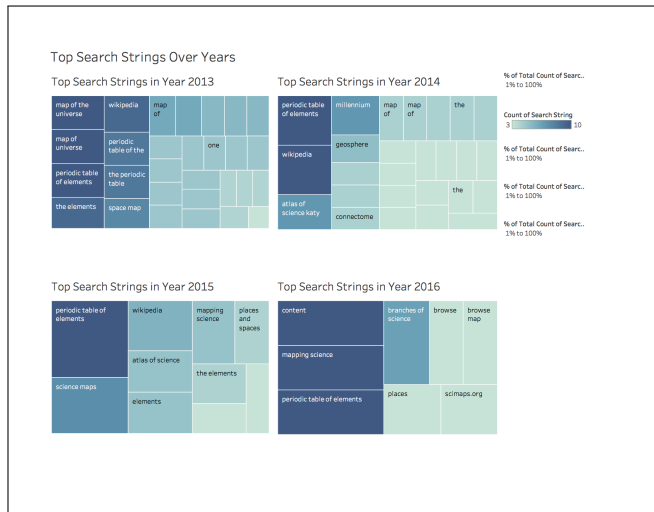
### 4.1 Top Countries and Trend

¡TODO: Sharad¿

### 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 websitefis 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.

### 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
- 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

**Figure 4: Top Search Strings in year 2013 to year 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 4 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 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 doesn't seem to have any effect of the website redesign as we observe a minor rise. A drastic change could have implied the user's 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 doesnt seem to have addressed that issue.
- The website redesign doesnt 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

¡TODO: Sharad¿

## 4.7 Dashboard

¡TODO: Leonard¿

## 5 KEY INSIGHTS

TODO: Leonard

## 6 OTHER SIMILAR TECHNOLOGIES

¡TODO: Shreyas¿

### 6.1 Google Analytics

### 6.2 Piwik

## 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

## REFERENCES

[AgastiAgasti2017] Avadhoot Agasti. 2017. ivmooc-scimap-webanalytics. (2017). https://github.com/avadhoot-agasti/ivmooc-scimap-webanalytics/ Online; accessed 03-Apr-2017.
[TABLEAUTABLEAU] TABLEAU. Tableau Desktop. (????). https://www.tableau.com/products/desktop Online; accessed 03-Apr-2017.
[WeblizerWeblizer2017] Weblizer. 2017. The Wenlizer: What is it. (2017). http://www.webalizer.org/ Online; accessed 03-Apr-2017.