

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 Rawagad

School of Informatics and Computing, Bloomington, IN
47408, U.S.A.

srawagad@indiana.edu

Sharad Ghule

School of Informatics and Computing, Bloomington, IN
47408, U.S.A.

sghule@indiana.edu

Leonard Mwangi

School of Informatics and Computing, Bloomington, IN
47408, U.S.A.

lmwangi@indiana.edu

ABSTRACT

TODO: Sharad

KEYWORDS

ACM proceedings, L^AT_EX, text tagging

1 INTRODUCTION

TODO: Sharad

Lorem Ipsum is simply dummy text of the printing and typesetting 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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, Washington, DC, USA

© YYYY ACM. 978-x-xxxx-xxxx-x/YY/MM...\$15.00

DOI: 10.1145/nnnnnnnn.nnnnnnn

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 Weblizer tool [WeblizerWeblizer2017]. Weblizer 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 queriable store: 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 support importing of 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 follow 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.

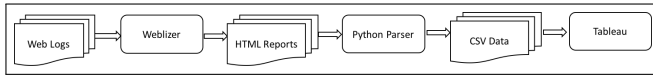


Figure 1: Data Pipeline.

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 Webanalyzer 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

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-bovack.pdf
5	1787	0.42%	3078	0.01% /home/panel
6	1536	0.36%	139776	0.28% /Scripts/query-1.11.1.min.js
7	1504	0.36%	73984	0.15% /Scripts/query_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.

Year	Month	Day	HitsTotal	HitsPct	FilesTotal	FilesPct	PagesTotal	PagesPct	VisitsTotal	VisitsPct	SitesTotal	SitesPct	KbytesTotal	KbytesPct
2007	3	1	6287	2.20%	5026	2.30%	5545	2.30%	855	2.84%	237	4.08%	227572	3.09%
2007	3	2	3944	1.31%	2889	1.33%	3075	1.31%	351	2.38%	225	4.09%	203449	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	1.33%	243	1.80%	184	3.35%	148816	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%	7597	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.

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.

3.3 Visualizing in Tableau

Explain how different reports are generated and then integrated in one dashboard

4 VISUALIZATION

TODO: Avadhoot: Just introduction of section - 2 liner;

4.1 Top Countries and Trend

TODO: Sharad;

4.2 Top Agents and Trend

TODO: Shreyas;

4.3 Top Referrals

TODO: Leonard;

4.4 Top Searches and Trend

TODO: Avadhoot;

4.5 Most Popular Pages

TODO: Shreyas

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 ACKNOWLEDGEMENTS

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

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.