[redacted]

[redacted]

Matt Ruppel

Spring 2018

Internship

Internship Research Paper

Introduction:

At E-rehab, we make websites for other clients (specifically, physical therapy clinics, but most of this research should apply to a wider audience). Part of the services we do for clients is marketing, and the better our clients do, the better we do as they refer other clients and boost our reputation. Thus, it is useful to research how we can improve our clients' websites and market effectively for them. Obviously, getting impressions from visitors/clients is directly useful, but it is very hard to get random visitors to leave feedback and clients may not actually know what their visitors are looking for. It is much more ideal to have quantitative data about visitor's habits and preferences, which is where analytics is often used. However, standard analytics systems are designed to only track one or a few websites individually, not determine trends among hundreds of websites. Besides finding improvement metrics, we were also looking for ways we could prove to customers our services are worth it. One of the ways we planned to do it was tracking when a visitor scheduled an appointment with a customer's clinic. The customization were were looking for led us to the decision to build our own specialized analytics system. This system would be able to aggregate data across many websites and give us presumably useful metrics about the websites. Therefore, the question becomes: How can specialized data collection and analytics improve the product web designers deliver to their clients?

Background:

E-Rehab is a private company offering website and marketing services to physical therapy clinics. The main difficulty E-rehab has in selling these services is convincing clients that websites will directly benefit their business—specifically, increasing the number of clients they have. Web tracking and analytics, defined as passively collecting and analyzing data from website visitors, can solve this problem along with providing crucial information about what can be done to improve websites. Using web analytics provides additional flexibility and credibility over methods such as surveys both in terms of raw data points and trustworthiness of the data sources themselves.

Web analytics is not a new concept, and many large companies exist that provide web data collection and analytics as a paid service. These services vary heavily in scope and purpose. There are also many free tools for website owners to set up their own analytics, such as Piwik and Google Analytics. Since E-rehab provides services to hundreds of clients which individually have one of more websites apiece, the most important feature in analytics is the ability to aggregate data from all websites together easily. Without aggregate data, it is near impossible to draw meaningful and general conclusions about the websites' designs. Additionally, being able to classify web pages (Burby and Angie) and aggregate "traffic" (webpage visits) by the type of page visited has been shown to be able to give useful insight for what website visitors are looking for (Norguet). However, none of the currently available web analytics tools support this use case. Therefore, a new analytics system was built to match these needs.

While researching how to design a tracking system, the biggest resource was analysis of already existing systems. For storing user's information, many used relational SQL databases

because of their benefits for linking different types of user data and profiles to individual traffic (Harrington). A SQL database is a type of digital data storage where data is stored in “tables” that are very similar to spreadsheets except the contents of each row/column must be a fixed, defined type. The SQL part of the name refers to the fact that the database accepts queries, or questions about the data, in the SQL programming language. The relational part refers to spreading data over multiple tables that are “related” by a common value, such as the visitor’s ID. Once data is stored in a SQL database, it can easily be statistically analysed (Pavlo).

Methods:

A tracking system was created to collect raw data points (Fig.1) into a SQL database about website visitors. Information was tracked about a visitor on their first page visit on any site tracked by the system, when they visited a page tracked by the system. The advantage of the database’s design is that data from all websites is already aggregated but easily separable. This allows analysis performed on the data to be applicable to all websites as a whole rather than specific websites.

Fig.1 - Example database rows for the three different types of data tracked. For definitions of terms, refer to Appendix B.

(a) Row generated on a visitor's first visit. Each row contains the visitor's unique ID (referred to here as the GUID, or globally unique identifier), user agent, IP address, and the website ID they first visited.

(b) Row generated on a page visit. Each row contains the visitor's GUID, visited website ID, page URL, page type, page return code, and an UNIX timestamp.

When a tracked website is visited for the first time by each visitor, the visitor is assigned a random unique ID. The ID is saved as a cookie (a small piece of data that is attached to a site and a visitor's browser) on the tracking system's website. Whenever a visitor then visits a page from then on, the system will use the cookie to identify the visitor and record data. Since each website may have a completely unique URL structure, the system will reference a database to determine the "page type" visited.

There was initially problems with getting the browser to contact the tracking system's server[1]. When the page normally tells the browser to contact a server, it also tries to let the page see the response from the server. However, if any website could normally get the contents of any page, it would be a huge security problem. Therefore, browsers don't let pages get the contents of pages on a different website unless the response contains a specific string allowing all or specific websites to access the response. This works normally, but the tracking system needed access to cookies, and for that, the page needs to make an "authenticated" request—and there is no way to let all websites, only specific lists, get the response of one. The technique used to get past this was to tell the browser to run the contents of the page's response. This technique is

[1] The page visit count not be tracked on initial page load for two reasons: the cookie with the user's ID is on another domain and therefore would not be sent with the request, and it would break the site's caching setup.

called cross-site scripting (XSS), and is normally used to exploit websites by running code from another page on them, but in this case it was used in a legitimate manner[2].

Once data has been entered into the database, analyses can be done with special "SQL queries" that can quickly process the large volumes of data and produce statistics that can be easily graphed. For example, the SQL query "SELECT * FROM traffic WHERE guid = 'abc'" would return all the data rows in the "traffic" table where the 'guid' (visitor ID) field equals 'abc'. For analytics, one of the queries used would be "SELECT page_type, count(1) FROM traffic GROUP BY page_type ORDER BY count(page_type) DESC". This returns the page type and count of each page type, ordered by the number of rows with the page type - descending. The data returned by the SQL query can then be fed into Google Charts, the base that powers Google Spreadsheet chart functions, to produce graphs like ones in the analysis section.

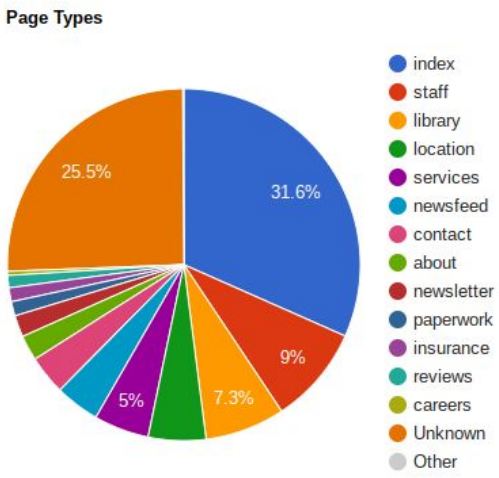
Research Timeline

Date Range	Completed Work
April 9-22	Tracking system research and structure design
April 23-29	Tracking system code
April 30-May 6th	Tracking system testing
May 7-10	Data collection and analysis

[2]The most common use of this is called JSONP.

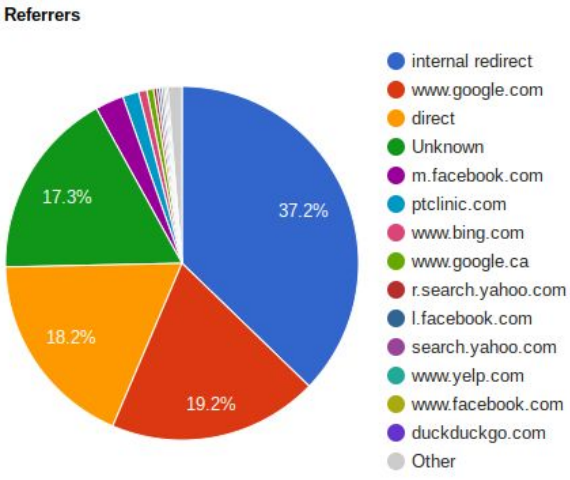
Analysis:

The first data analyzed was the most visited page types (Graph A). (It should first be noted that all data includes visits from robots.) Unsurprisingly, the index (main page) was the most popular page type, having more than a three times share of total traffic compared to the second highest page type. Interestingly, pages about the company's staff ranked second in popularity, over the pages describing the locations of places themselves or "about us" pages. This indicates that visitors are often interested in this information, therefore focus should be put on making information about staff more clear and thorough, possibly additionally including it on the main index. However, the "about" pages may be ranked less because that information is on the front page and therefore easily available. Additionally, over 25% of page visits were to pages that did not have an assigned category (mostly because of inconsistent naming schemes between websites). Many of these pages are actually "location" pages, accounting for the low ranking of the "location" type, but it was not realistic to assign each unique page a category within the time frame of the research.



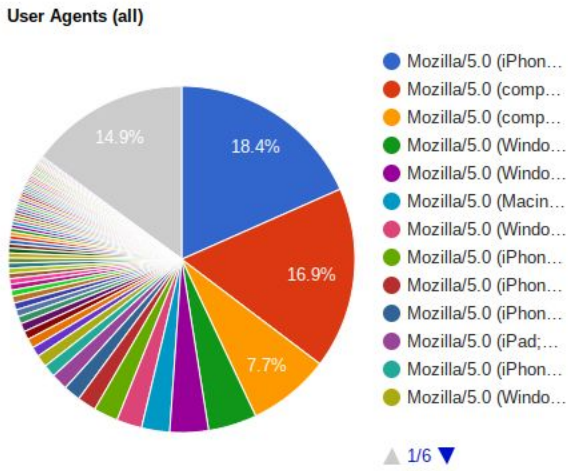
Graph A

Next, the most common referrers to each page visit were analyzed (Graph B), grouped by the referring domain (base website). Internal redirect indicates that a page was referred from a link on the same website, such as going from the main page to the "About Us" page. In other words, over 35% of traffic was from retained visitors. This is a ratio of 1:2 to traffic from external sources. The most popular external referrer was Google, compromising almost 20% of traffic - at least 5 times more than other search engines combined. This justifies the focus on optimizing websites to appear high on Google searches, and matches the results of previous analyses (Pozadzides). Direct visits (the visitor directly entering the website's URL into the address bar) compromised just under the amount that visits from Google did, indicating a lot of visitors hear about the site from a direct source, such as visiting the clinic in person. Facebook was the next significant individual referrer, which makes sense considering websites are marketed on Facebook with ads. It is interesting to note that many visitors had an "unknown" referrer. The researchers were unable to figure out what the significant of these are, or how they came to be.



Graph B

Part of the data collected once for each visitor was the visitor's user agent (browser/device identifier). There are a massive amount of unique user agents, so the data is very spread out, but some trends can be seen. The most popular user agent was for Safari on iOS (Apple) 11.3 (Piejko), followed Google and Bing's web robots. Data about the most current most common user agents across the whole web does not appear to exist, but this shows that mobile devices actually compose a majority of traffic the researched websites receive, highlighting the importance of optimizing and designing websites to display well on mobile devices. It also indicates just how much of web traffic comes from robots from big search engines — over 25% of traffic.



Graph C.

Conclusion:

The most statistically significant result of current data analysis is how much of the traffic comes from Google and Facebook. This gives justification for focusing on marketing towards those sites and indicates we should this. Google's market share also potentially means that it may

not be worth it to invest into optimizing for other search engines like Bing and Yahoo. The next significant trend is the prevalence of mobile devices, and how the most common user agent was a mobile Apple device by a very significant margin. As even Google has said themselves, more and more of web traffic is coming from mobile devices, and Google now even prioritizes mobile sites over their desktop versions (Zhang). This implies we should focus heavily, possibly even more so, on mobile versions of websites as opposed to their desktop versions. Although the common page type analysis was interesting, it is not clear how it is immediately useful especially considering the number of uncategorized page visits.

Following this research, work should be done to make sure as many page visits are accurately categorized as possible. It would also be beneficial to perform a large-scale analysis of all user agents so that it can be determined what percentage of traffic is from mobile versus desktop visitors, and even how many visitors are bots. Data in this research was not separated from bot traffic, and as such the results may differ if that data is removed. It is highly recommended that that is done for future analysis. Outside of websites for physical therapy clinics, these results especially relating to user agents, referrers, and mobile traffic should be applicable to all website developers, and the page categorization approach useful for those who operate many sites.

[redacted]

Works Cited

Burby, Jason, and Angie Brown. "Web Analytics Definitions." *Web Analytics Association*, WAA Standards Committee, 16 Aug. 2007, www.digitalanalyticsassociation.org/Files/PDF_standards/WebAnalyticsDefinitionsVol1.pdf.

Harrington, Jan L. *Relational Database Design and Implementation, 4th Edition*. Morgan Kaufmann/Elsevier, 2016.

Norguet, Jean-Pierre, et al. "A Page-Classification Approach to Web Usage Semantic Analysis." *Engineering Letters*, vol. 14, no. 1, 12 Feb. 2007, www.engineeringletters.com/issues_v14/issue_1/EL_14_1_21.pdf.

Pavlo, Andrew, et al. "A Comparison of Approaches to Large-Scale Data Analysis." *Proceedings of the 35th SIGMOD International Conference on Management of Data - SIGMOD 09*, 2 July 2009, pp. 165–178., doi:10.1145/1559845.1559865.

Pozadzides, John. "Analysis: What Are the Web's Top Sources of Referral Traffic?" *Readwrite*, 28 July 2010, readwrite.com/2010/07/28/analysis_what_are_the_webs_top_sources_of_referral_traffic/ .

Piejko, Pawel. "List of User Agent Strings." *DeviceAtlas*, 7 Mar. 2018, deviceatlas.com/blog/list-of-user-agent-strings.

Zhang, Fan. "Rolling out Mobile-First Indexing." *Google Webmaster Central Blog*, Google, 26

Mar. 2018, webmasters.googleblog.com/2018/03/rolling-out-mobile-first-indexing.html.

[redacted]

Appendix:

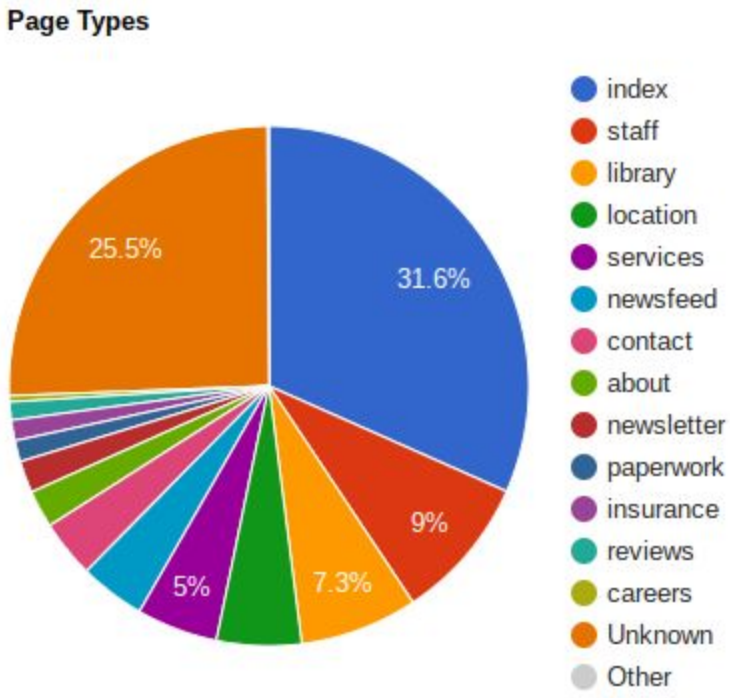
All graphs created with the free and permissively licensed Google Charts.

Interactive versions of these charts statically hosted at https://sera-sch.github.io/paper_sp2018/.

Check them out!

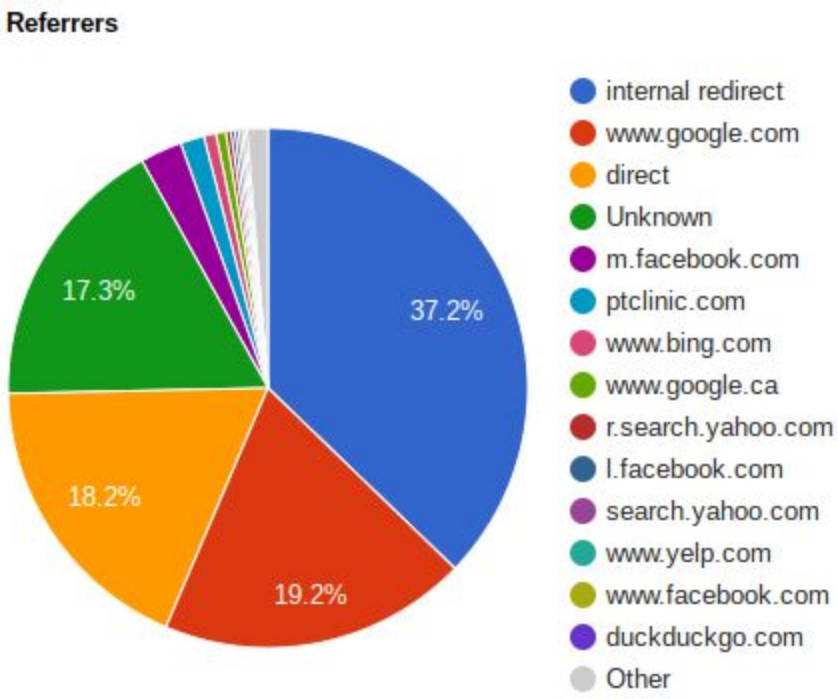
Graph A: Most popular page types, pie chart

Query: `SELECT page_type, count(1) as c FROM sc_ws_traffic GROUP BY page_type ORDER BY c DESC`



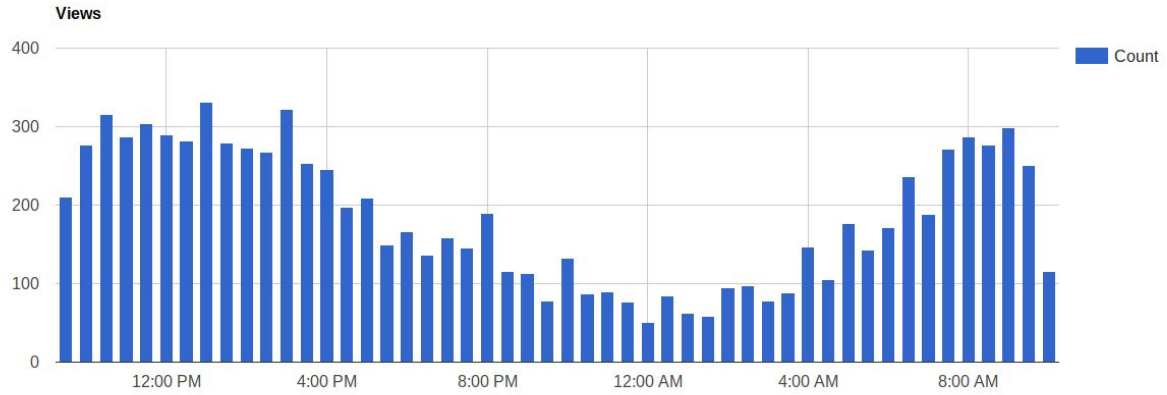
Graph B: Most common referral domains, pie chart

Query: `SELECT referral_domain, count(1) as c FROM sc_ws_traffic GROUP BY referral_domain ORDER BY c DESC`



Graph C: View data per hour for G5 sites, per half hour, column chart

Query: `SELECT COUNT(1) as c, DATE(FROM_UNIXTIME(timestamp)) as d, HOUR(FROM_UNIXTIME(timestamp)) as h, FLOOR(timestamp/(60*30)) % 2 as tk FROM sc_ws_traffic GROUP BY d, h, tk`



Graph D: Most common user agents, pie chart

Query: `SELECT user_agent, count(user_agent) as c FROM sc_ws_cookie GROUP BY user_agent ORDER BY c DESC`

Note that the gray upper left quadrant is grouping all the user agents that compose individually under 0.1% of visits.

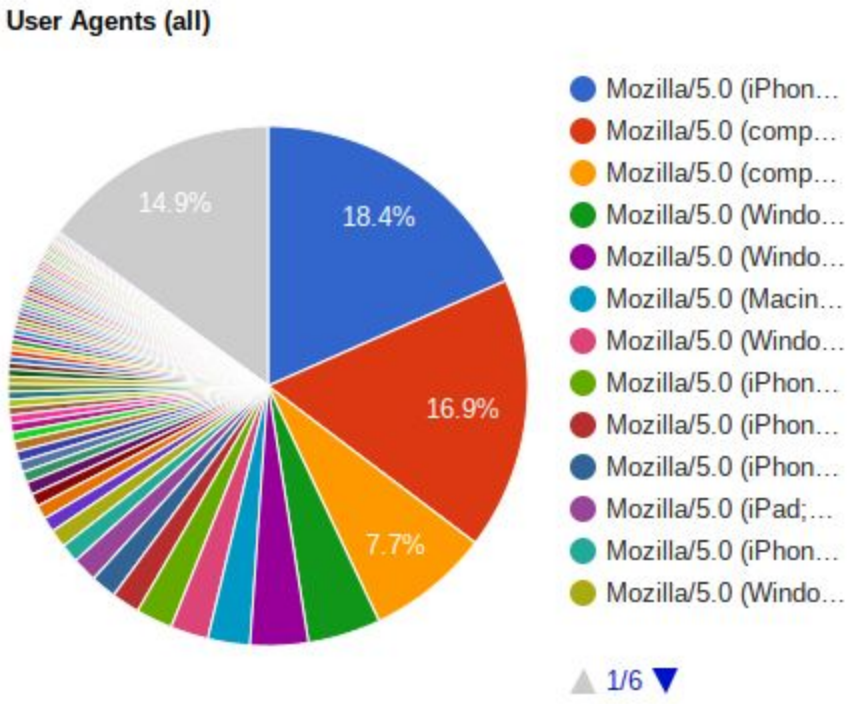


Table:

User Agent	Count/%	User Agent	Count/%
Mozilla/5.0 (iPhone; CPU iPhone OS 11_3 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/11.0 Mobile/15E148 Safari/604.1	18.4% (911)	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/66.0.3359.139 Safari/537.36	4.5% (225)
Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)	16.9% (840)	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/534+ (KHTML, like Gecko) BingPreview/1.0b	3.6% (178)
Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)	7.7% (380)	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_4) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/11.1 Safari/605.1.15	2.6% (130)

Appendix A: About Data Time

The system time zone of the MySQL database is EST (PST+3). All dates are converted to PST for display.

Data time range: May 10 9:30 ~ May 11 10:00

[redacted]

Appendix B: Definitions

Cookie: A small piece of data attached to a site and browser

Database:

Database row: A single, literal row of data in a two dimensional database with defined structure.

Event: An significant, arbitrary action done while on a page by a visitor.

IP address: A string that uniquely identifies each router on the internet.

Return code: A number a website sends the browser, sometimes called a status code. All the name alludes to, the number represents the state of the server: 200 means "OK", 404 means "page not found".

SQL database: A database that is able to accept SQL queries to process its data.

SQL query: A string written in the SQL programming language that tells the database which data to output and how to process it, if at all.

UNIX timestamp: A number counting the seconds that have passed since 1/1 1970 00:00 UTC.

User agent: A string that identifies a web browser and device type. Should (but not guaranteed to) be unique across different web browsers, browser versions, and device types.

Examples:

```
Mozilla/5.0 (iPhone; CPU iPhone OS 11_3 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/11.0 Mobile/15E148 Safari/604.1
```

iPhone 11.3 running Safari 11 (build 604.1)

```
Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)
```

Google's automatic website analyser, version 2.1