Visualizing website traffic

Fernando Gómez-Herrera[a], Raúl Monroy Borja[a], Dante Dessavre[b], José Ramírez Márquez[b]

[a] *Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM)*
[b] *Stevens Institute of Technology (Stevens)*

Abstract

Digital advertising has become one of the most popular business model for many companies REF. Thus, the need of tools for analyzing and measuring digital content has become an important area of interest in past few decades. Nowadays there are common standards for measuring digital content, from web site traffic (page views, bounce rate, session count) to advertising campaigns effectiveness (conversion rate, funnel techniques). Most commonly the use of text reports are used for presenting such analytics. The present work provides a tool for analyzing graphically the interactions of users in web sites, proposing new layouts for displaying visits and enabling powerful detections like browsing patterns, profiling objective pages or even spot visitor clusters.

Keywords: web analytics traffic-analysis

1. Introduction

Web analytics has become an important area of research in the past few years REF. With companies building their business models based on advertising, it is very important to have tools for measuring digital content interactions such as number of page views, session count, demographics and other commonly defined metrics like *bounce rate, exit rate, conversions, etc.*, using such tools you can get to know better your audience and so what are their preferences, what is the content they like, how much time spend reading or view content and also the browsing behavior inside your website. Big companies like Google are aware of such need of tools in the market and because that advertising is their main business model, they have been creating tools for their customers to analyze digital metrics. The most popular services provided by Google are Google Analytics and DoubleClick. Both tools can be used to measure web advertising campaigns and they provide predefined reports for that. Of course Google is not the only one in the advertising market, Yahoo, Microsoft, Amazon and even Facebook have their own tools and services for their platforms.

Email addresses: gomezhyuuga@acm.org (Fernando Gómez-Herrera), raulm@itesm.mx (Raúl Monroy Borja), dgamades@stevens.edu (Dante Dessavre), jmarquez@stevens.edu (José Ramírez Márquez)

2. Previous work

There are many tools available for measuring digital content and because this is such an important topic in the *marketing* industry, some companies have created their own tracking tools and build a business around it. We will analyze such tools dividing them into two categories: *enterprise solutions* and *open source* tools. Many of the enterprise solutions (with the exception of Google products) are paid services whereas the open source tools are free to use.

2.1. Enterprise solutions

Google Services. As the biggest company in the Internet, Google has three main products focused on web advertising (Google AdWords and DoubleClick) and web analytics (Google Analytics). Combining such products you can have powerful insights of the users that are visiting your website. DoubleClick is a paid service, but AdWords and Analytics are free to use.

Google Analytics is the main product for getting reports and analyze the traffic to a website. It can be configured to import and track ad campaigns from AdWords and DoubleClick, and allows you to segment the traffic from many sources and with many filters as you want. Also it has the advantage for being widely known by marketing experts and people in many areas so it has become a standard in a certain way.

comScore. It is an American company founded in Virginia which now has big presence not only in the Internet but also providing services to other kind of media companies like the TV industry, newspapers, health care and others as well. Nevertheless, we can not get a further analysis of their tools because they are paid services and usually oriented to big companies with large amount of data. Despite of this fact, comScore has been very open with their current research and publishing some reports in a periodic way [1, 2].

KissMetrics. Also a paid service but this company focuses in more tailored need experience, focusing on consulting and teaching their customers how to implement the tool and interpret the results. Popular stories about the use of KissMetrics are the one of LucidChart (https://www.kissmetrics.com/lucidchart/), a tool for creating digital diagrams, which after upgrading their product, needed to measure the performance of the new design and by using KissMetrics, they could determine an increase of 30% more conversions. Another case is the one of the e-commerce site Manillo, an Amazon-like service in Denmark, which increased their ROI by 50% by understanding better their audience (https://www.kissmetrics.com/manillo/).

The previous tools can only be tested as a hosted service. This is a disadvantage for some companies that need to complain with law regulations about the storing the customer's data or for companies that need an on-premise solution. Another big concern about this kind of services is that usually you do not own the data, instead you only have access to it through a third-party.

2.2. Open Source tools

Piwik. One of the most popular and robust tools that could be self-hosted or as a service in the cloud. Piwik is a company that focuses on giving their users a complete control of everything, meaning that you get full reports (no data sampling), you are the owner of the 100% of your data and also it is completely open source, which means that you can customize it as you wish. It is developed in PHP and also provides an HTTP API for consulting the reports.

Open Web Analytics (OWA). Although this project has not published any new version since 2014, it is still popular in legacy websites. It was integrated in former versions of Content Management Systems like Wordpress or Media Wiki. It can be tested only by installing it on a personal server. One of the greatest features included out of the box is the *click heatmap* that shows the *hottest* sections of your pages.

2.3. Common concepts

The following terms are commonly used for referring to the performance of a website or an advertising campaign. It must be mentioned that each service provide their own implementation for computing the values and it is very common for the tools to disagree with the results.

- Page View. Number of times a page was viewed by an user. A user can view the same o many pages multiple times in each visit.
- Visit. A collection of events from a user in a given period of time. A visit can contain multiple page views, downloads, clicks or any custom event triggered. Usually platforms setup a *idle time* to finish the visit, i.e. if the user did not make any action within a window of time (e.g. 30 mins), then his/her visit is *closed.* After that, any action would be considered as a new visit. The terms *visit* or *session* are used interchangeably within platform services.
- Bounce Rate. This is commonly used as a Key Performance Indicator (KPI) of how your pages are performing. It indicates the number of users that only view one page and leave immediately. The measure is more meaningful when used for single pages or goals, instead of using it in a global way.
- Objective/Goal. Custom defined event that gives to your company value or even revenue. Common objectives are: page views of selected pages (cart, about us, offers); for e-commerce adding an item to the cart, making a purchase, the checkout page; for media sites playing a video, watching a video for at least X seconds, etc. Each company define their objectives according to their business and needs.
- Conversion. Amount of users that triggered or *achieved* an objective. For example how many times an objective page was uniquely visited, how many purchases have been made, number of checkouts, times an advertisement was seen, etc. Each conversion normally has associated a revenue for the company so increasing the conversions is a KPI for any analyst.

A complete description of other terms used in media analytics can be found in a report made by the Web Analytics Association (waa) [3].

As an example of how the platforms present the previous metrics, Figure 1 shows examples of some of their reports.

3. Motivation

Getting insights from the data is a very important skill for any business analyst, marketer or any decision maker. However, it is almost impossible to get knowledge from *raw data*, so it must be presented in a way that is easy to understand and analyze. Luckily, nowadays there are many options to visualize data; as we mentioned in Section 2 some of those platforms provide tools for doing such tasks, and by the use of geographic maps, pivot tables, heatmaps and other visualization techniques, spotting valuable information becomes easier. Nevertheless, the trend of creating new ways to display web analytics has not been changing much in the past few years, objective reports, conversions and site performance usually are still displayed as tables or big score counters. Figures XYZ shows an example of those reports.

This research propose new ways to display website traffic by using an interactive tool that provides several ways to arrange visits, conversions, user behavior, click stream, pages and filtering options which combined with, for example, Machine Learning pattern recognition, could spot clusters for new market niches or common visitor segments.

4. Methodology

The main objective of this research was to use other ways to display visitors and their information in a more graphically way, rather than using tables and counters. To this end, we experimented some techniques such as *treemaps*, *heatmaps* and *network graphs*. Due to the nature of the Internet being a network of connected computers, we decided to base us in network graphs. Following this decision, we needed a support library to compute and display network graphs in many layouts and one of the most popular available is Cytoscape, which was developed by the University of Toronto []. There are another alternatives like D3.js [] or even Sigma.js [] but Cytoscape has a more friendly programming interface and also it includes a powerful feature to make *queries*. Cytoscape provides a bunch of graph theory algorithms such as *Dijkstra shortest path* [], *aStar search [], MST algorithms [], BFS/DFS traversing and even PageRank [] and centrality algorithms [].*

Taking advantage of this library we build a web application written in ECMAScript 6 and as a front-end framework, we chose Facebook's React [].

4.1. Data collection

In order to collect the information from visitors, we mounted a website and used one of the tools previously mentioned in Section 2: Piwik. This was because it allowed us to gather a lot more of raw information and so we could manipulate it to our needs, things that we could not make if we would use for example Google Analytics. Also there was the problem of getting traffic; we generated *fake traffic* by using HTTP botnets [] but there were also some days of human traffic to the website.

5. Results

The motivation of this research was not to build something that magically at first sight you could spot the information you are searching for. Instead, we think that each business, each scenario, each website provides information in a unique way and because of that each *analysis process* is different. Following this philosophy, the analyst just needs the tools and he/she will make use the most of it to find what is important. The work proposed in this research is intended for that; the tool is there, now use it and find useful information. In order to accomplish this task, of course, the tool must provide useful features, and the one proposed in this work has some interesting ones. Lets take a loot to some of them.

5.1. Main interface

The user interface is split in two main areas: the network graph on the left side, and the details pane on the right side. All the information is displayed given the date range selected in the top bar; there is also the possibility to change that time period.

5.2. Data representation

Visits. Each visit is represented an *orange* circle.

Pages. Represented as *blue* circles.

Objectives. For pages marked as *goals or objectives* there is a different color to highlight them, so the *green* circles are used.

Figure XYZ shows the circle's style and how they are connected between them. Each *visit node* is connected to one or more *page/objective nodes* an thus, the connection represents a single visit.

The tool aims to provide easy ways to represent and style the nodes. For example you would be able to chose the size of the circle depending on a metric. For example, making bigger page nodes depending on the number of page views it receives. The same principle could be applied to visits by the use of metrics like *visit duration, visit count, number of actions made, etc.* And for the case of objective pages a good indicator would be the *bounce rate.*

5.3. Click Path

This feature provides a detailed view of how a specific user is browsing inside the website. It is displayed again as a graph but using a DAG (Directed Acyclic Graph) [] layout. Such graph starts with an orange node (the visit) and it should end in a page node. Intuitively in that path there should be at least one or more green circles, representing that the user reached a goal and so making a conversion for the business.

This is a powerful view of how users are navigating in the website, contrasted with tables or lists which are used by some tools (Figure XYZ), by using a graph we can spot more easily the complexity of the visit. This could be useful for example to detect behavior. As it was mentioned in Section 4, there was traffic generated by bots and humans in this research and the *ClickPath* feature can spot clearly the differences between the navigation complexity. As you can see in Figure XYZ, human traffic shows a more complex navigation than bots.

5.4. Segmentation

One of the first tasks when you start analyzing an audience is to create segments, which is no more than performing a filter process given a series of conditions; e.g. young people (`a < age < b`), men living in New York City (`gender = 'male' && location = 'NYC'`), people coming from social networks (`referrer IN ['Facebook', 'Twitter', 'Snapchat']`), etc. So it is very important to have an easy way to perform such filters and because of that, there is a *querying* console incorporated in the tool. This work is presenting only the user interface to access the already built-in Cytoscape's querying system. The following are examples of queries you can make using the tool, supporting the base logic operators: *AND, OR, NOT, relational operators* ($=, >, <, \leq, \geq$)*, string matching* and others. The full capabilities are provided by Cytoscape and the full specification can be found in the documentation []. The aim of this section is to provide examples of how to use such query system to our advantage to create segments.

5.4.1. Query format

Depending on how your data is structured and the available attributes, you would be able to perform queries matching such attributes. We are using data coming from Piwik so we are able to query visits and pages based on their database structure []. The format follows the next grammar:

$$group[attr\ OP\ value]$$

Where:

$group \in node, edge, className$

$className$ represent a custom class assigned to a particular data. In our case we have three classes: `visit, page, objective`. So instead of using *node or edge*, we use such classes which is less abstract.

attr means the attribute that you are using as a filter. Such attributes should correspond to the context represented by the *group.* So for example, for the *visits* group, you could have attributes like `duration, country, browser, events, etc.`; on other hand, in the case of the *page* group you could have attributes like `url, visits, pageviews, bounceRate, exitRate, avgTimeSpent, etc`.

OP represents any binary operator. Depending of the *data type* you would use one of another. The available operators are: `=, !=, >, <, >=, <=, *=, ∧=, $=`. The last three operators are used for string comparison.

value is the value used for matching *attr* by the selected operator (*OP*). Depending of the *data type* of *attr* you would put strings, numbers or booleans.

5.4.2. Logical operators

To implement the logical operators *AND, OR* you have to follow the next format:

AND:

$$group[attr_1\ OP_1\ value_1]group_2[attr_2\ OP_2\ value_2]\ \ldots\ [attr_n\ OP_n\ value_n]$$

Notice the join of groups after the square brackets.

OR:

$$group[attr_1\ OP\ value_1], group_2[attr_2\ OP\ value_2]$$

Notice the use of the *comma (,)* for concatenating conditions.

5.4.3. Examples and use cases

Visitors from the United States:

```
.visit[countryCode = "us"]
```

Visitors using Google Chrome *and* with a visit duration greater than 10 seconds *and* that are from South America.

```
.visit[browserCode = 'CH'][visitDuration > 10][continentCode = 'ams']
```

An interesting use case for the previous querying system enables us to use it with **Machine Learning** algorithms, specifically with decision trees based techniques that create patterns splitting the data by attribute values. The figure XYZ shows a J48 decision tree [] trained with traffic data, spotting clusters of visitors that match certain properties.

5.5. Insight through layouts

Positioning the nodes in a clever way could result into instant insights from your site and how it is performing. By default Cytoscape provides several layouts to arrange the node, some of them would position the nodes depending on the graph's attributes like the node's *degree*, connected components or betweenness centrality. Default layouts included in the library are: *grid, random, preset, circle, concentric, breadth first, cose.* It also has extensions to use third-party layouts like Cose and Cose-Bilkent [] (often used in microbiology), Cola [] or force layouts [] which simulate gravity attraction. It is the decision of the analyzer to use the layout that seems to fit better to the situation, choosing the right one could improve the way to discover new patterns.

However, we propose the use of the *concentric layout* to give an indicator of one of the most used metrics in web analytics: *bounce rate.* The motivation is to *keep in the closest radius your objective pages*, meaning a good performance the more of they are in the first concentric circle. This enables the analyst to start asking the right questions. The following cases are possible situations of *performance indicators* like *bounce rate*, we believe that using the *concentric layout* would be a good way to spot such performance and also to discover new opportunities in the website pages. The metric used for the reports is *bounce rate.*

Case 1: green nodes in the first level.

This is a good indicator representing that your objectives are *healthy*, meaning that your are accomplishing all of your objectives, because all of them are in the first concentric level. Of course this would be the *ideal case* for any business.

Case 2: green nodes spread in many levels.

Another way to spot how your objectives are performing. The intuition here is that if the green nodes are far from the first level, then it means that you should improve them. It also serves as an indicator to *track performance*: for example, you can define *how many levels your objective has jumped since previous auditing.*

Case 3: blue and green nodes mixed in the closest levels.

Although at first sight you can think that it is not a good indicator to have *blue circles* inside your first level, *you can take advantage of it.* Because this is telling you that you have pages that can lead to *potential conversions*, therefore, you should try to use them to engage your users. Possible use cases for those pages are inserting in them call-to-action elements, displaying ads, making offers or treat them as landing pages. Again, you can use this knowledge to engage the users and lead them to a conversion.

6. Conclusions

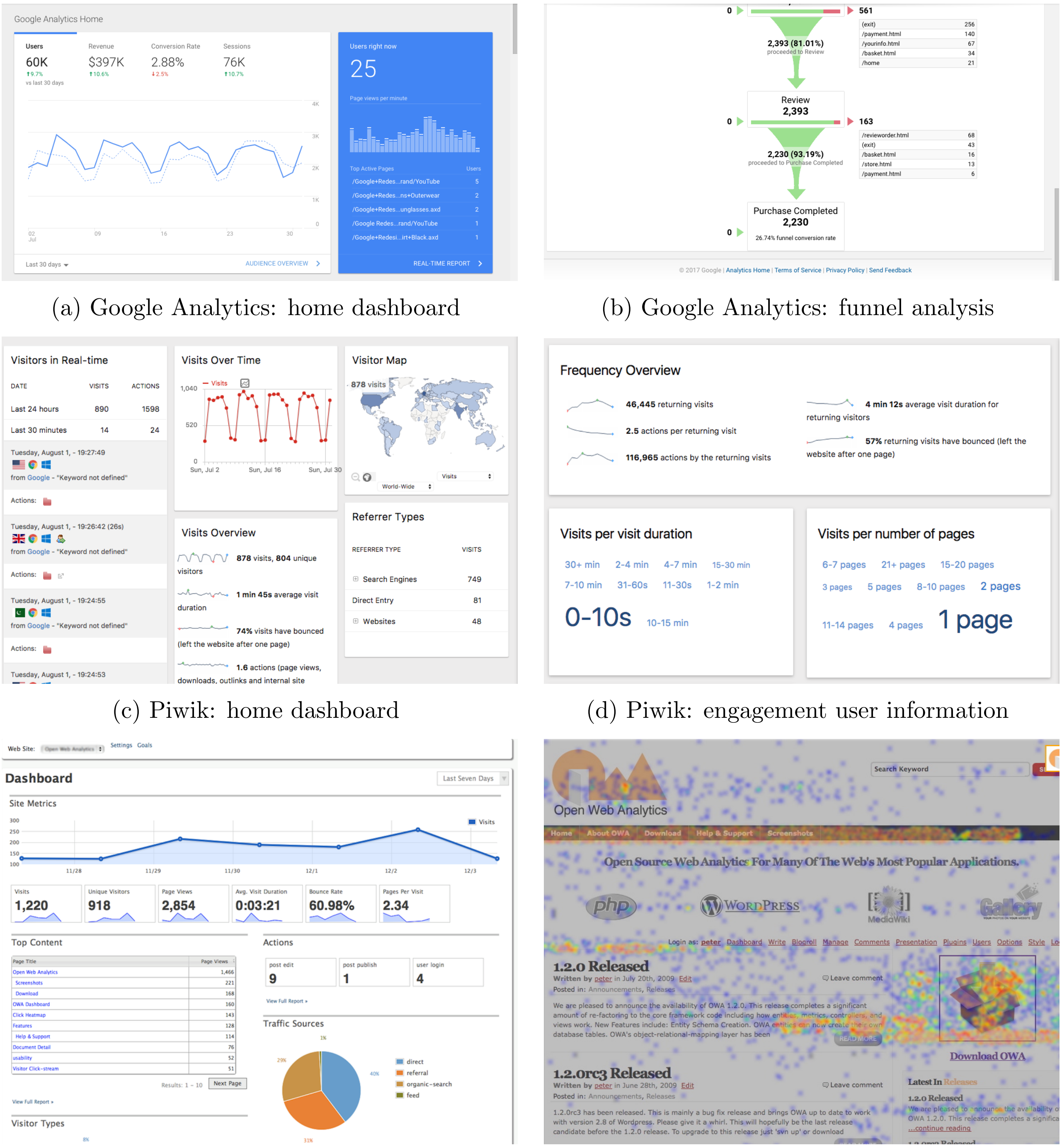
The work presented in this research proposes new ways to display website traffic which we believe offer an easy way to get insights of your performance more quickly.

Future work absolutely is related to Machine Learning techniques combined with the query system. There is great opportunity to match this two areas, for example some ways to do it is the creation of recommendation systems to suggest new audience segments automatically. Another use case could be recognizing user traffic as invalid (bot) or invalid (human) through the use of a classifier. Also, the tool could be improved in many ways, for example by the implementation of a auto-complete system for the query console, integration of new layouts, applying node/edge compression techniques and aggregation algorithms or saving common queries. We believe this tool provide powerful features to analysts and we will continue the development of this research, adding new features and proposing new visualization mechanisms.

[1] Invalid Traffic (2016).
URL `http://www.comscore.com/Products/Advertising-Analytics/Invalid-Traffic`

[2] Brian Pugh, Battling Bots: comScore's Ongoing Efforts to Detect and Remove Non-Human Traffic (nov 2012).
URL `http://www.comscore.com/esl/Insights/Blog/Battling-Bots-comScores-Ongoing-Efforts-to-Detect-an`

[3] Waa, Web Analytics Definitions [Report], Web Anal. Assoc. (4.0) (2007) 1–32.


Google Analytics Home
Users 60K ↑9.7% vs last 30 days
Revenue $397K ↑10.6%
Conversion Rate 2.88% ↓2.5%
Sessions 76K ↑10.7%
Last 30 days
AUDIENCE OVERVIEW
Users right now 25
Page views per minute
Top Active Pages
REAL-TIME REPORT
2,393 (81.01%) proceeded to Review
Review 2,393
2,230 (93.19%) proceeded to Purchase Completed
Purchase Completed 2,230
26.74% funnel conversion rate
561
163
© 2017 Google | Analytics Home | Terms of Service | Privacy Policy | Send Feedback
Visitors in Real-time
Last 24 hours 890 1598
Last 30 minutes 14 24
Visits Over Time
Visitor Map
878 visits
Visits Overview
878 visits, 804 unique visitors
1 min 45s average visit duration
74% visits have bounced (left the website after one page)
1.6 actions (page views, downloads, outlinks and internal site
Referrer Types
Search Engines 749
Direct Entry 81
Websites 48
Frequency Overview
46,445 returning visits
2.5 actions per returning visit
116,965 actions by the returning visits
4 min 12s average visit duration for returning visitors
57% returning visits have bounced (left the website after one page)
Visits per visit duration
0-10s
Visits per number of pages
1 page
Dashboard
Site Metrics
Visits 1,220
Unique Visitors 918
Page Views 2,854
Avg. Visit Duration 0:03:21
Bounce Rate 60.98%
Pages Per Visit 2.34
Top Content
Actions
post edit 9
post publish 1
user login 4
Traffic Sources
Visitor Types
Open Web Analytics
Open Source Web Analytics For Many Of The Web's Most Popular Applications.
1.2.0 Released
1.2.0rc3 Released
Download OWA

(a) Google Analytics: home dashboard

(b) Google Analytics: funnel analysis

(c) Piwik: home dashboard

(d) Piwik: engagement user information

(e) OWA: home dashboard

(f) OWA: click heatmap

Figure 1: Dashboards and reports of some analytics platforms