Online Advertising: Impact Analysis & Future Direction

Anagha Lagu & Bhavya Ghai

Abstract

Online Advertising is critical for existence of millions of Websites. Ad-Blockers have impacted the business model of numerous websites drastically. There's an ongoing struggle between Websites & Ad-Blockers. Ad-Blockers try to block all possible Ads and websites try to sneak their ads through or force users to disable ad blockers. We need to find a middle ground where websites can sustain themselves and end-users are exposed to less intrusive ads that provide a more positive user experience. In this project, we are trying to solve this problem by devising an intelligent ad-blocking technique. We have analyzed website content and have assigned them a rating based on their ad content. The rating is assigned on a scale of 10 based on parameters like number of ads, download time, download size & screen space occupied by ads on the respective websites. The websites with rating above a certain threshold have been classified as white-listed sites and won't be affected by Ad-blockers. This novel technique will encourage ad-servers & content providers towards Acceptable ads which will be a win-win situation for everyone.

Introduction

Few things are more frustrating than when you notice your website loading slowly. There is a myriad of factors that contribute to performance bottlenecks and slow page speeds. One of the notorious culprit can be ad networks. Although a large number of websites sustain on ads, its effect has not been studied much. Ads can have a significant impact on page load time, battery life, user privacy and overall online experience. Since Online advertising is a critical economic driver which funds wide variety of online services, it's important to study its impact on page load time of different web pages.

Identifying Ads

Over 90% of the ads we consume are provided by ad servers. An ad server stores information about ads and delivers them to one or more web sites for display to visitors. Ad servers also track ad displays, clicks on ads, and generate statistical reports. For this project, we are using a list of popular ad servers. We'll analyze website content and look for urls which belong to our list of as servers. Later, we'll classify them as acceptable or unacceptable

Acceptable Ads

Acceptable ads are a set of non-intrusive ads which ensure positive online experience. They can be defined as a set of ads which obey a set of constraints with respect to number, size, placement, etc. of ads. They help publishers, networks and advertisers generate more viable sources of revenue, which keeps online content free and delivers less intrusive ads that provide a more positive user experience.

Problem Statement

The research project involves:

1. Identifying ad content on webpages – number of ads, download time for each ad resource, download size for each ad resource, screen space occupied by ads.
2. Analyzing the effect of ads on page load time of webpages
    - Comparative study for webpages belonging to different categories – news, social, sports, knowledge, shopping etc.
    - Comparative study on various platforms and browsers.
3. Using machine learning techniques to classify websites as acceptable and non-acceptable by devising a novel website rating generation technique based on analyzed ad content of websites.

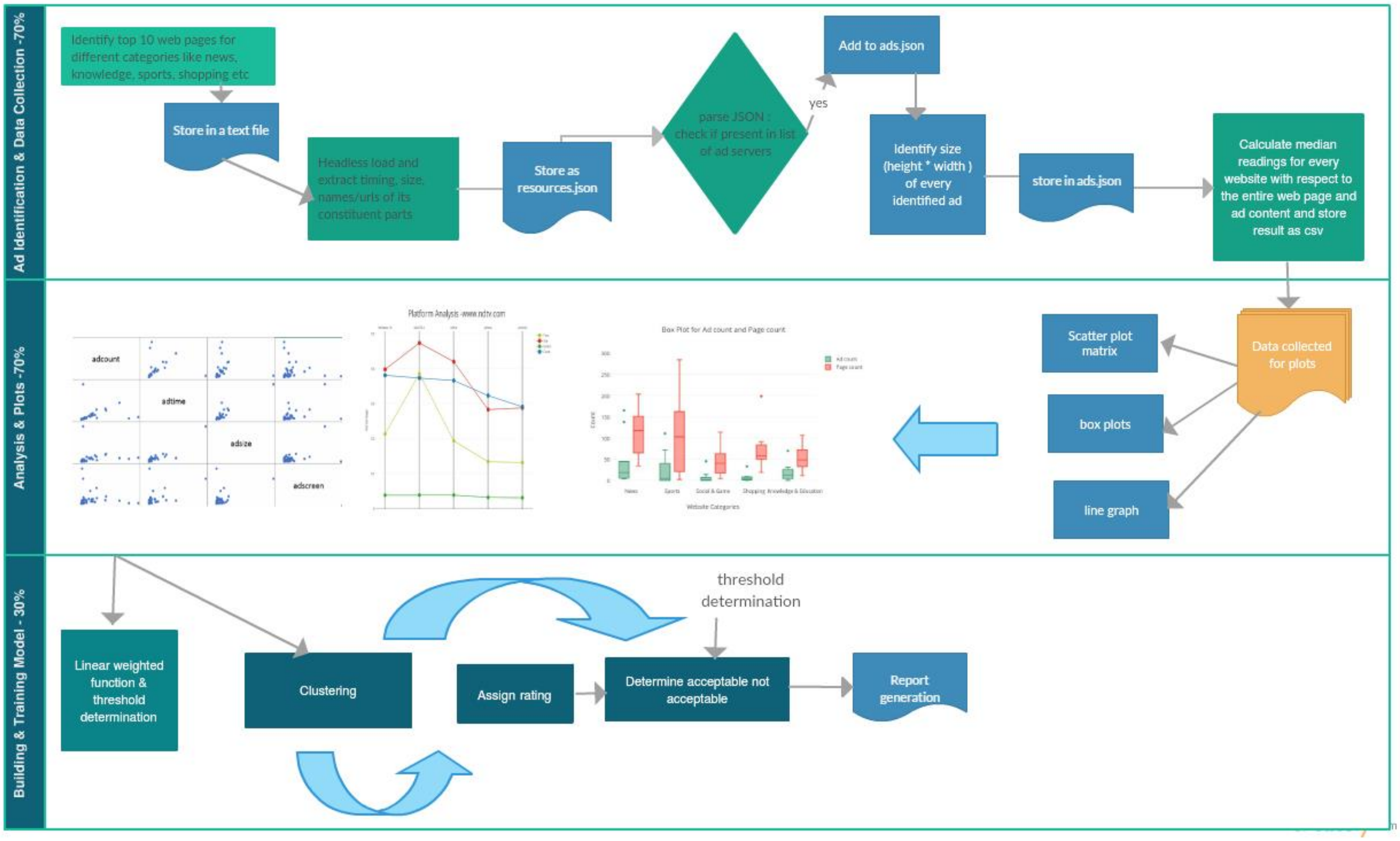
Methodology

Performance testing the effect of ads on page load time of a web application is tricky since it involves accurately identifying ad content objects in a website and measuring the time and space required by each ad object individually.

We have come up with a system using JavaScript, Python and shell scripts that would accurately identify and measure the aforementioned things, in an effort to understand how page load time was affected due to ads under various conditions of the experiment.

The system involves three components namely web page parser, ad object detector and data and analysis report generator for various metrics under consideration. The web page parser uses a JavaScript library called PhantomJS to headlessly test page performance. The parser loads the web page under consideration and logs the size, time and details of its constituent parts. It also lists the fastest and slowest resources along with the largest and smallest. Further the generated logs are used by the ad object detector to segregate the ad content from the non-ad constituent parts. Typically advertisement URL's have certain keywords like "google.adsense" etc. by which they can be uniquely identified. Such keywords are used by the ad identifier to lookup in the list of ad servers. If they match, the URL is labelled as an advertisement. Further the ad object detector creates a JSON file containing identified ads with their respective load time and size. Further the screen dimensions of the identified ads are determined using JavaScript code and they are appended to the JSON file. This exercise is done 6 times and a median value of the four parameters namely ad count, screen size, download size and download time is determined for every ad. The analysis report generator uses this data to generate graphs and charts for various metrics under consideration. Besides this, the system also maintains a config file that decides how it should run. For example, you can set the user-agent header of the request made by PhantomJS to request the page, in case you're serving mobile apps off similar entry-point URLs to your desktop content.

Workflow


Ad Identification & Data Collection -70%
Identify top 10 web pages for different categories like news, knowledge, sports, shopping etc
Store in a text file
Headless load and extract timing, size, names/urls of its constituent parts
Store as resources.json
parse JSON : check if present in list of ad servers
yes
Add to ads.json
Identify size (height * width) of every identified ad
store in ads.json
Calculate median readings for every website with respect to the entire web page and ad content and store result as csv
Analysis & Plots -70%
adcount
adtime
adsize
adscreen
Platform Analysis -www.ndtv.com
Box Plot for Ad count and Page count
Ad count
Page count
Count
News
Sports
Social & Game
Shopping
Knowledge & Education
Website Categories
Data collected for plots
Scatter plot matrix
box plots
line graph
Building & Training Model - 30%
Linear weighted function & threshold determination
Clustering
Assign rating
threshold determination
Determine acceptable not acceptable
Report generation

Tools and Technologies

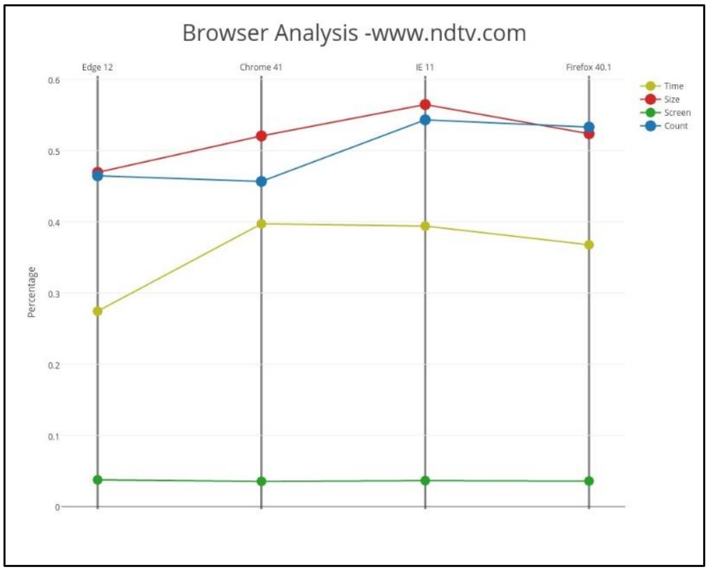
1. Languages
   - Python
   - Shell scripts
   - JavaScript
2. Libraries
   - Phantomjs
   - Plotly
   - Matplotlib
3. Tools
   - Excel

Evaluation Setup

To analyze the effect of ads on page load time of webpages a comparative study was carried out for websites belonging to different categories like news, social, sports, knowledge, shopping etc. and on various platforms like iPad, mobile and desktop and browsers like Firefox, chrome etc.

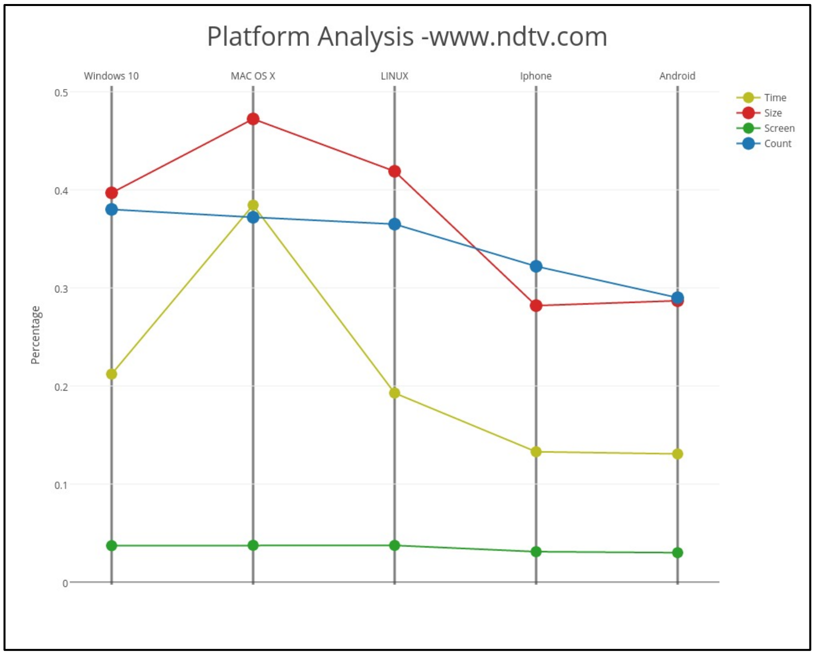
Various graphs were plotted and measurements were done considering 4 metric parameters like download time, download size, ad count/ number of ads and ad dimensions. Interesting trends were observed and they have been summarized below.

Sample Results – Plots



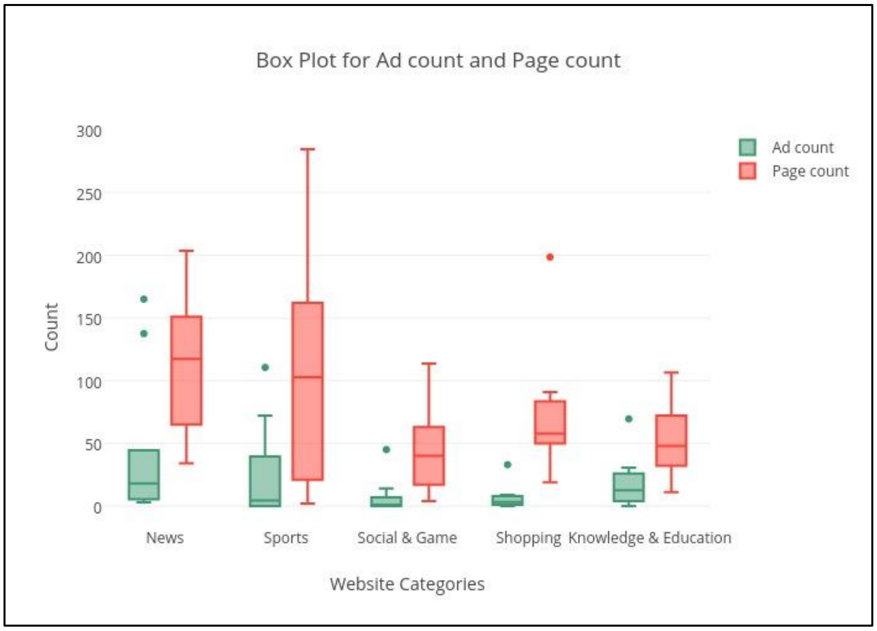
The website ndtv.com was loaded on Mozilla, Chrome, Internet Explorer, and Edge browsers. Four measurements namely ad count, ad download time, ad download size and ad screen space were evaluated and a parallel co-ordinate graph was plotted with the browser categories on x-axis and the percentage of the four metrics with respect to the page count, page download time, page download size and page screen space on y axis. Interesting trends have been observed.

1. The screen space remains constant for all browsers.
2. Since browsers render resources differently, the ad download time and download size varies on different browsers.
3. Similarly, the ad content i.e. the number of ads varies with respect to browsers.

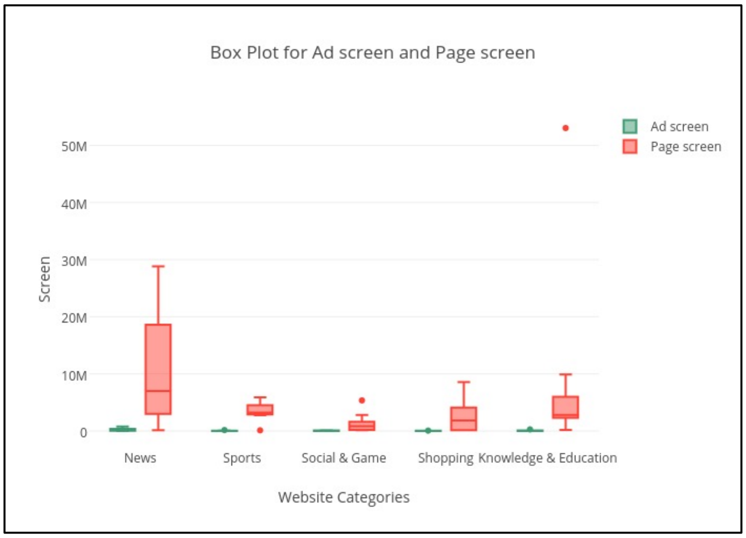


The website ndtv.com was loaded on Windows 10, Mac OS X, LINUX, IPhone and Android. Four measurements namely ad count, ad download time, ad download size and ad screen space were evaluated and a parallel co-ordinate graph was plotted with the platform categories on x-axis and the percentage of the four metrics with respect to the page count, page download time, page download size and page screen space on y axis. Interesting trends have been observed.

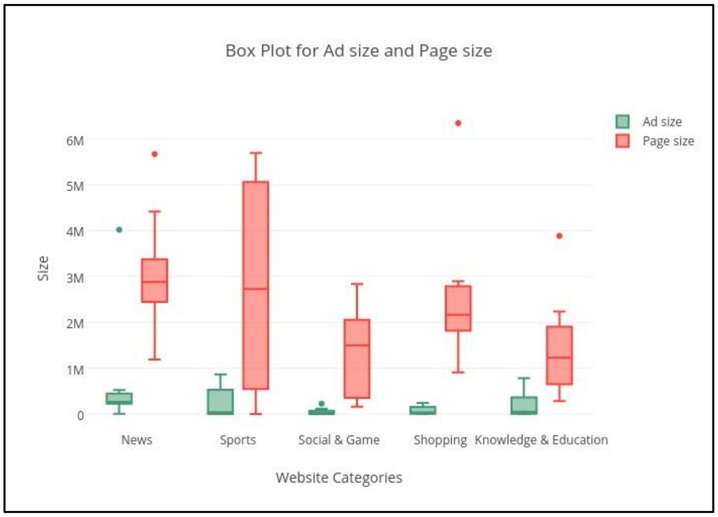
1. Since the available screen bandwidth changes for desktops and mobile devices, the amount of ad content displayed varies.
2. Thus, due to the variation in ad count, the download time and download size varies.
3. The percentage of screen occupied by ads remains proportionate to the total page screen size since the number of ads varies.

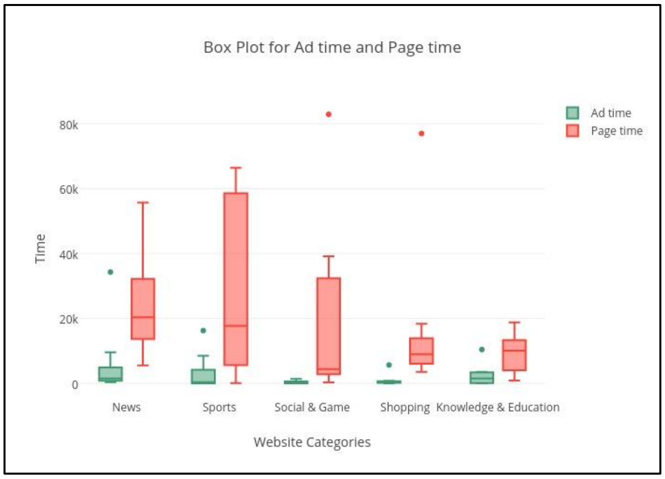


The plot indicates that websites under sports and news categories have more ad content than websites under other categories like shopping, knowledge, and education and social and games.

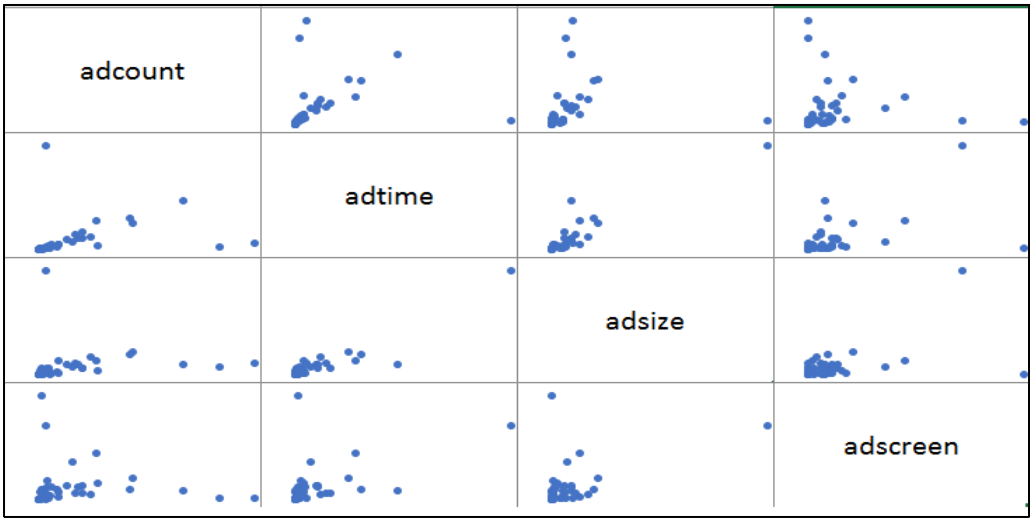


Ads occupy a large amount of area on news websites. Although the ad content was more for categories like news and sports, shopping websites had few ads occupying a larger screen space.





The ad download time and download size was measured under a constant bandwidth. There was a direct correlation between the number of ads and the download time and download space. Although social and games websites had fewer ads than news websites, the type of ad content they hosted was different. Thus we could conclude that the type of ads also determine the download time and download size.



The scatterplot matrix tries to see trends in correlation patterns among all possible combinations of attributes.

Future of Online Advertising -- (30% Portion)

Ad-Blockers :- Ad-blockers have no functionality on its own. It doesn't block anything until we "tell" it what to block by adding external filter lists. Filter lists are essentially an extensive set of rules that tell Adblockers which elements of a website to block. Filter list are maintained by an online community of users who write filters when they encounter a new ad. We can add any filter list we like. For example, we can block tracking and/or malware. We can also create your own filter lists. Almost all filters are open source, therefore many filter lists have been created by Internet users.

Filter lists enabled by default include:

- An ad blocking list selected based on language (EasyList)
- The Acceptable Ads list

EasyList corresponds to our browser language and is aimed at disabling ads that are considered to be intrusive by our community of users. The Acceptable Ads list displays ads that comply with the acceptable ads criteria agreed upon by our community of users.

Acceptable Ads :- Acceptable Ads help publishers, networks and advertisers generate more viable sources of revenue, which keeps online content free and delivers less intrusive ads that provide a more positive user experience.

Proposed Solution

- Find middle ground :- Acceptable Ads
- Unblock Acceptable Ads so that sites can sustain
- Block non-acceptable ads which ensures good online experience
- Rate websites based on Ad content
- Classify websites based on some threshold value
- Store the result in some centralized server
- Ad-blockers access centralized server while loading website
- Update server contents after some fixed interval

Ranking Techniques

We'll try to rank websites based on the ad content. We will be using four parameters :- Ad count, Screen space occupied by ads, Ads size & Ads download time. We will calculate the ratio of these resources with the resources required for the entire page and calculate ranking. We'll calculate ranking on a scale of 1-10 with 10 being the highest rating. The site with highest rating will have no ads. The site with rating of 1 will have only ads. Below, I have listed two ranking techniques.

Linear Weighted Ranking

In this technique, we'll assign weights to each of the parameters. Then we can simply calculate the ranking by computing the sum of the product of weight and parameter value. Thereafter we will scale the rating on a scale of 1 – 10 and choose some threshold value. Lastly, we will classify websites based on the rating value.

Clustering

In this technique, we'll consider a 4D cube of unit side. The point (0,0,0,0) will correspond to the ideal case with no ads and the point (1,1,1,1) will correspond to the worst case with only ads. We will plot all websites as a point in this 4D space. The two extreme points as two cluster centers. We'll then try to classify websites based on the distance of website from either cluster center as shown in the following diagram.

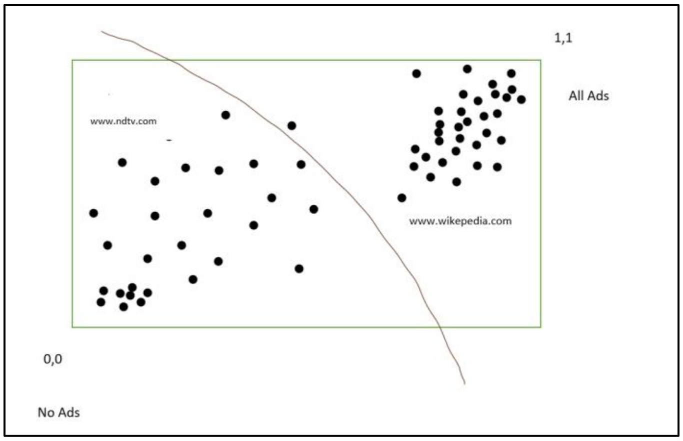
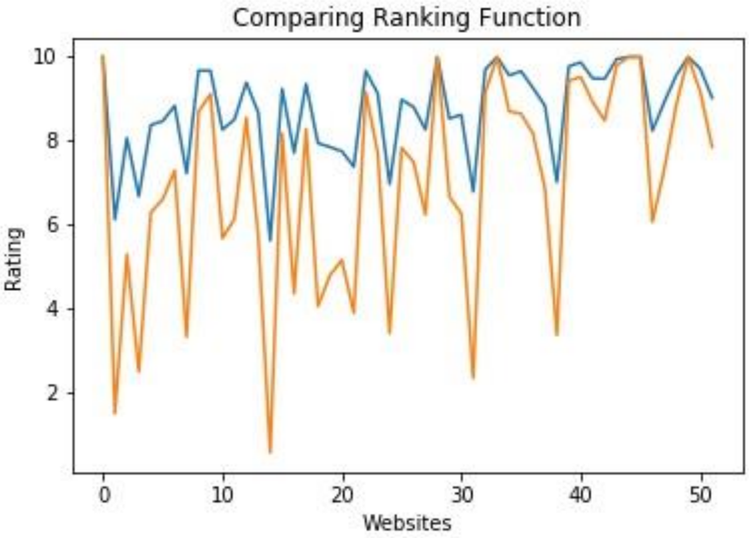


Figure 1 Clustering Technique to rate websites

Next we tried to compare both ranking systems as shown in the following diagram. Here blue line shows the linear ranking rating and orange line represent clustering rating. As we can see, both look similar overall with some minor differences. We can look into results for more description. There is no ground truth to validate ranking systems so we can use crowdsourcing in future to validate ranking systems.



Conclusion

In this project, we have tried to analyze the Impact of Online Advertising on end users. We extracted different parameters to characterize ad content of a website. We tried to compare ads for various platforms, browsers, website category, etc. & tried to unearth interesting patterns. Additionally, we worked on the problem of Ad-Blockers. We suggested a novel rating technique to classify websites based on their ad content. Such transparent rating system will prompt different websites to enhance their rating and prevent from getting their ads blocked.

Future Work

Future work may involve crowdsourcing for calculating the default threshold value for different ranking Systems. Crowdsourcing can help us quantify the extent to which ads can be shown without much affecting the end user experience. We can also look to add additional parameters to define acceptable ads. We can try to identify ads location on screen, classify ad images into flashy/non-flashy, etc.

Git Repository Link :- Code Link

https://anagha_lagu@bitbucket.org/anagha_lagu/fcn_project.git

References

[1] Allowing acceptable ads in Adblock Plus. Retrieved from https://adblockplus.org/acceptable-ads [Accessed May'17].

[2] Easylist. Retrieved from https://easylist.to/ [Accessed May'17]

[3] PhantomJS. Retrieved from http://phantomjs.org/ [Accessed May'17]

[4] Orr, Caitlin R., et al. "An approach for identifying JavaScript-loaded advertisements through static program analysis." *Proceedings of the 2012 ACM workshop on Privacy in the electronic society*. ACM, 2012.

[5] Ad server hostnames. Retrieved from https://pgl.yoyo.org/as/ [Accessed May'17]

[6] Vattikonda, Bhanu C., et al. "Empirical Analysis of Search Advertising Strategies." *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*. ACM, 2015.

[7] Vattikonda, Bhanu C., et al. "Interpreting advertiser intent in sponsored search." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

[8] Hidayat, Ariya. "Phantomjs." *Computer software. PhantomJS. Vers* 1.7 (2013).

[9] Friesel, Rob. *PhantomJS Cookbook*. Packt Publishing Ltd, 2014.

[10] Ghose, Anindya, and Vilma Todri. "Towards a Digital Attribution Model: Measuring the Impact of Display Advertising on Online Consumer Behavior." (2015).

[11] Guha, Saikat, and Srikanth Kandula. "Act for affordable data care." *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*. ACM, 2012.