For this project I will be visiting a websites internal pages and comparing the data gotten from that to the results gotten from visiting a sites landing page.

Motivation

The original Princeton study "Online Tracking:A 1-million-site Measurement and Analysis " looked at one million pages. A problem with this is that they only visited the landing page and did not go into the internal pages the paper recognizes this as a fault but does not do it. When a user goes to a website they are not expected to stay only on the landing page for their entire visit and usually end up going to internal pages. By visiting the internal pages we can mimic more closely what a user might do and get a closer measurement to what a user experiences.

Problem

By visiting the internal pages of a website will we see a difference in the amount of http requests that internal page requires when compared to the landing page. If we observe a difference we can then claim that only looking at the landing page provides a skewed view on how websites behave, since users will tend to visit the internal pages.

Approach

For this project I will be using BeautifulSoup and OpenWPM. I will be using BeautifulSoup to find the internal pages and then open them in openWPM. I will be visiting 5 randomly chosen internal pages per site and taking the average of requests then comparing that to the main page. I am choosing to look at 5 internal pages since an internal page can vary in the content it could provide. We could end up choosing an about page and getting a small amount of requests or we could choose an article and get a lot of requests so by doing 5 pages we will get a closer glimpse to what the average amount of requests for a internal pages.

Related Work

Research has already been conducted for this problem but my search will differ with the way I search for internal pages. A research paper that has already done this is called "On Landing and Internal Web Pages" in that paper they explained how they got the internal pages from the Google Search Engine. While doing the top results from an engine is a good way I want to see if there any differences in results with my approach in finding internal pages between my search for internal pages and theirs. The paper also found good reason for including internal pages in studies to provide a clearer view of the internet.

Analysis

I will be graphing the amount of request from landing pages and the average from internal pages and seeing if there is a significant difference between them. I will then look for any significant difference between how many third-party requests internal pages makes versus how many landing pages make. I will also compare the types of third-party requesters to see if there are any differences.

Running my Program

In order to reproduce what I did you would have to run Getlink.py which gets all the internal pages from the websites it can visit.

AllSubPages.txt contains all the internal pages we found.

SubPageLess.txt cotains 5 randomly chosen internal pages from what we found.

NoSubPageGet.txt contains all the websites that we were unable to get internal pages from.

From there you would have to run SearchSitesOpenWpm.py in order to traverse all the sites. This program requires SubPageLess.txt for the list of internal pages it will visit

From there run AnalyzeData.py and that will show you two graphs one for the amount of third-party request and another for who the third parties are.

AnalyzeData.py also includes the method used to compare my list to the HisPar list

Search For Internal Pages

DownSides

A downside I noticed pretty soon is that many websites either did not re-

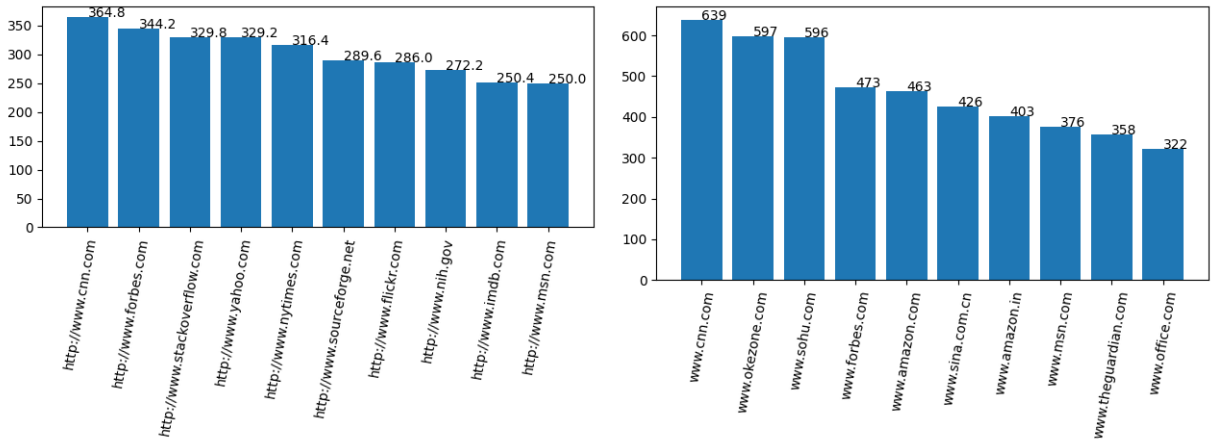
spond to my program trying to get the internal pages or the html I got from my scraping did not contain. This is due to many websites not wanting to be bombarded with requests from people scraping their website. I did a manual scrape on some of the websites to see what I was getting and some of them returned an almost blank html file. This led to my list of websites being used for a search being drastically reduced from 100 to 43. This was a huge reduction in websites gotten and makes this method have much less reach than the method used to get the HisPar list.

UpSides

An upside was in the types of internal pages we got from our search compared to HisPar list. My method got very different results compared to the google search results and in some cases I argue that my method might be better. A website where I think my method gives internal pages that users are more likely to click is in yahoo.com. For the hispar list the yahoo.com results are categories within the news such as sports, tech or finance, while my method returned article links on the main page. I believe that some users will mainly click articles in the main page so including them in a study would be helpful. Another example is cnn.com the hispar list had links to a bunch of coupon internal pages while my list had links to the subcategories of news such as sports, tech and finance. I also checked to see how many internal pages both methods found. I did this by inserting the internal pages from hisList into a list in python and checked to see if any of the sites I found where in the HisList. My comparison told me that they were 0 similarities between both list. Which I did not expect. This showed to me that although the HisPar list is better at visiting all the websites it does not contain all internal pages and might miss some important ones such as news articles.

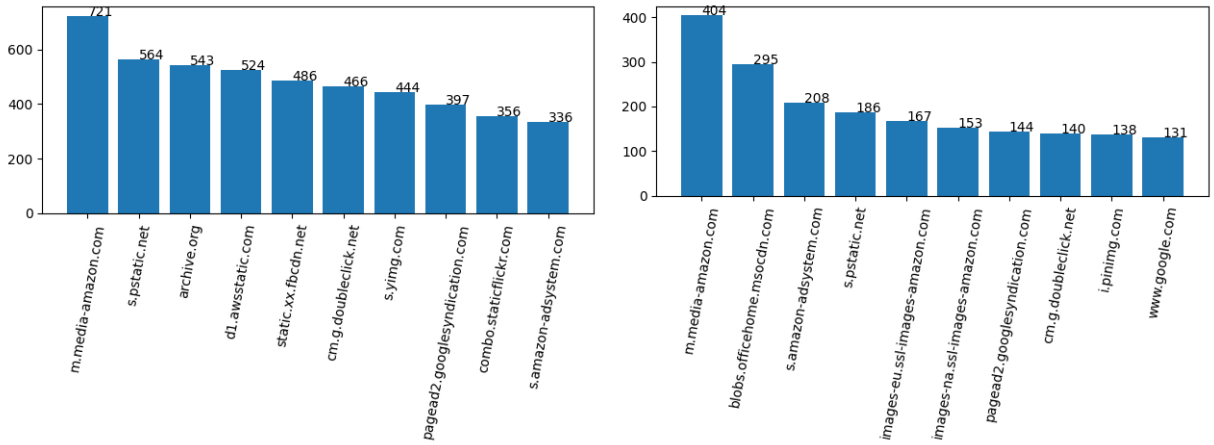
HTTP Requests

Although I was not able to get subpages for all 100 landing pages I decided to still graph my results to see if I found anything interesting.



Left is the internal page search. Right is landing page search.

I noticed a significant reduction in the amount of request made and that the websites that appear in the top 10. The difference in websites that appeared could be due to certain websites not accepting not returning 5 internal internal pages so they were ignored. While I believe the reason for a reduction in request especially in cnn.com is due to the type of internal pages I randomly chose. I noticed in particular with cnn I got internal pages that had /style in their url which are internal pages made to look more minimalistic.



Left is the internal page search. Right is landing page search.

Then with the type of third parties requested they stayed mostly the same. The numbers do not provide a fair comparison due to both not being able to visit all 100 pages in order to find the internal pages and I left the results from my internal search as is and did not average it out. However there is not that big of a change in who were the most requested

Conclusion

I did notice a difference in amount of http requests from websites specifically a reduction in requests. However, since I only chose 5 internal pages from a list might of skewed that a bit. Also not being able to get internal pages from the top 100 websites limited my comparison a significant amount. Though I do think this helps the idea that internal pages should also be used when looking at how prevelant third parties are. In my specific search I found a decrease in third party requests and although it might not be representative of all internal pages I think including them will still provide a better image of the internet. Also due to the types of third-parties websites are requesting from on internal pages that leads to me believeing that users are more or less equally tracked on internal pages and main pages

I did notice a difference in amount of http requests from websites specifically a reduction in requests. However, since I only chose 5 internal pages from a list might of skewed that a bit. Also not being able to get internal pages from the top 100 websites limited my comparison a significant amount. Though I do think this helps the idea that internal pages should also be used when looking at how prevalent third parties are. In my specific search I found a decrease in third party requests and although it is not representative of all internal pages I think including them will still provide a better image of the internet. Also due to the types of third-parties websites are requesting from on internal pages that leads to me believing that users are more or less equally tracked on internal pages and main pages