Analysis of Web syslog

Donthi Reddy, Sai Narsi Reddy

UMKC ID: 16186610

Email: [sdhy7@mail.umkc.edu](mailto:sdhy7@mail.umkc.edu)

1. **Introduction:**

In this paper, we present the daily analytics of networkrouting.net website between July-16th to August-19th of 2016. For total of 35 days, we present most visited webpages, daily 404 response code and daily chapter-1.pdf downloads. Along with these, we also present unique visitors and perform forecasting techniques on all the analytics.

Following, paper is arranged as follows: section - 2 provides the daily analytics for most visited pages, daily 404 response code, daily chapter-1.pdf downloads and unique visitors. Section-3 contains Forecasting results for all the section-2 analytics.

1. **Analysis:**
   1. **Most Visited Website:**

Here we find top-3 most visited webpages by finding all the unique webpages throughout the days, then taking the average visits for each webpage through days and sorting them in descending order.

Figure-1 shows the top-3 most visited webpages, with days on x-axis and daily visits on y-axis. the top three webpages are Index.html, robots.txt, and first chapter of Network Routing book.

The Index.html is the most visited page and it shows more jumps at the second half of the graph which means as the school days will get close more students will check the website most probably. The second most visited webpage is chapter-1 with more jumps at the second half of the graph which that also may be the school dates get close and students want to decide to take a look at the book to take this course or even study. Robots.txt is a text file placed in the web server and tells webcrawlers such as Googlebot if they can access a file. The robots.txt is among the top three visited pages and has almost a steady trend compared to two other graph which means the search engine constantly utilizing the robots.txt file to update the search engine server.

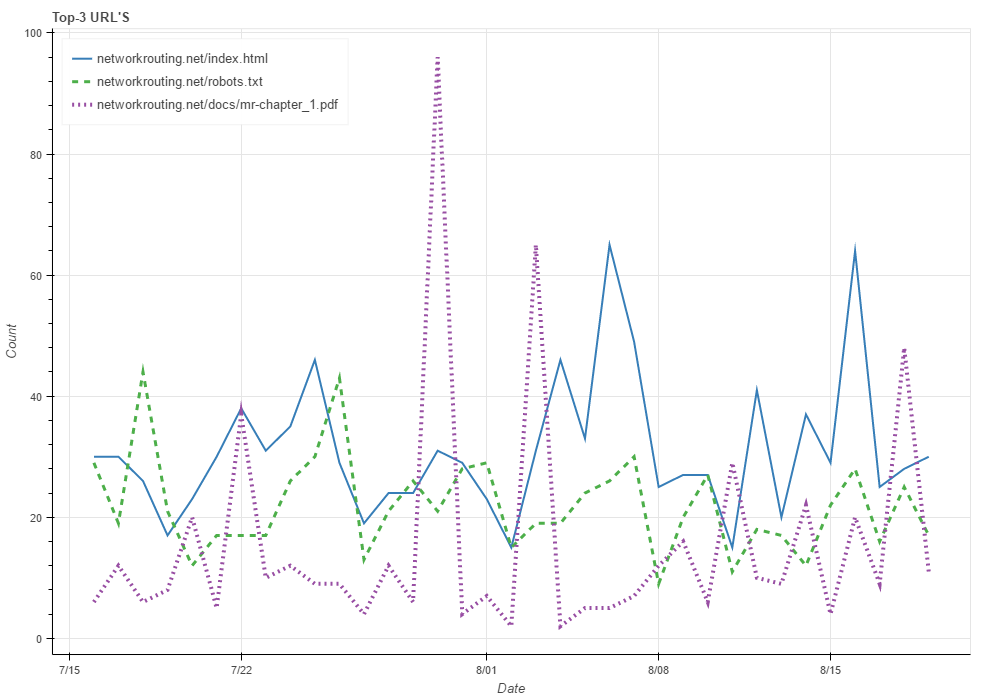


Figure 1 Top 3 URL's through days

* 1. **Daily 404 Count:**

The 404 is a HTTP response code indicating the client can communicate with server, but the server cannot find what is requested. Figure 2 shows the number of 404 code responses through all days.

At the second half we have more number of 404 that that is because the more requests for URLs are at those days and when webcrawler try to access pages like robots.txt, as there is no robots.txt page, more number of 404 responses happen.

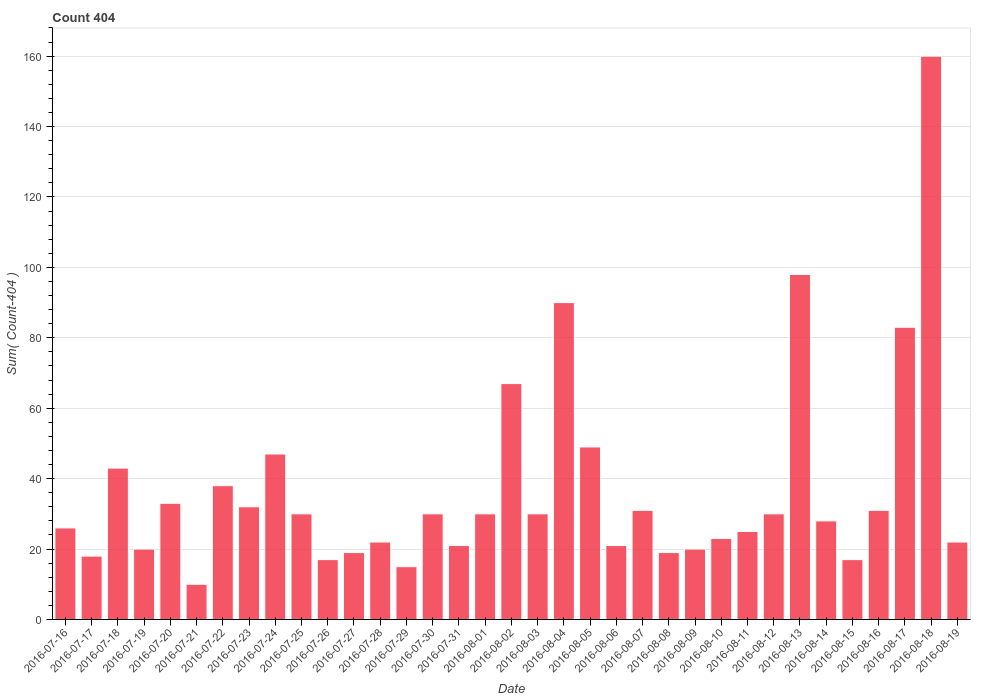


Figure 2 Daily 404 code responses

* 1. **Daily Chapter-1.pdf Downloads:**

From the figure-3 it can be seen that the number of chapter-1.pdf downloads is higher in month of July with 100 downloads as the maximum number of downloads in a day on July-30th 2016. In month of August, though the average number of downloads is low, there is a consistent number of downloads throughout the month.

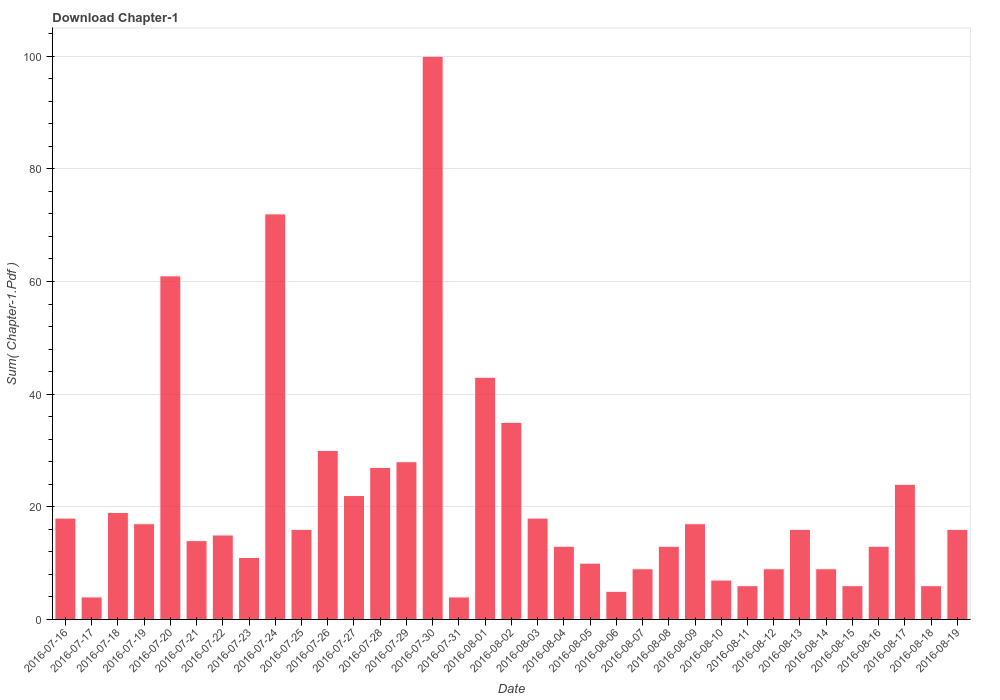


Figure 3 Daily Chapter-1.pdf downloads

* 1. **Unique Visitors:**

Unique IP addresses for each day are counted and plotted them as shown in the figure-4. It can be seen that, during month of august, at the mid-week there are more visitors and during the week-ends the trend goes down.

From the plot it can also be seen that the average visitor count is about 56 throughout the analyzed days.

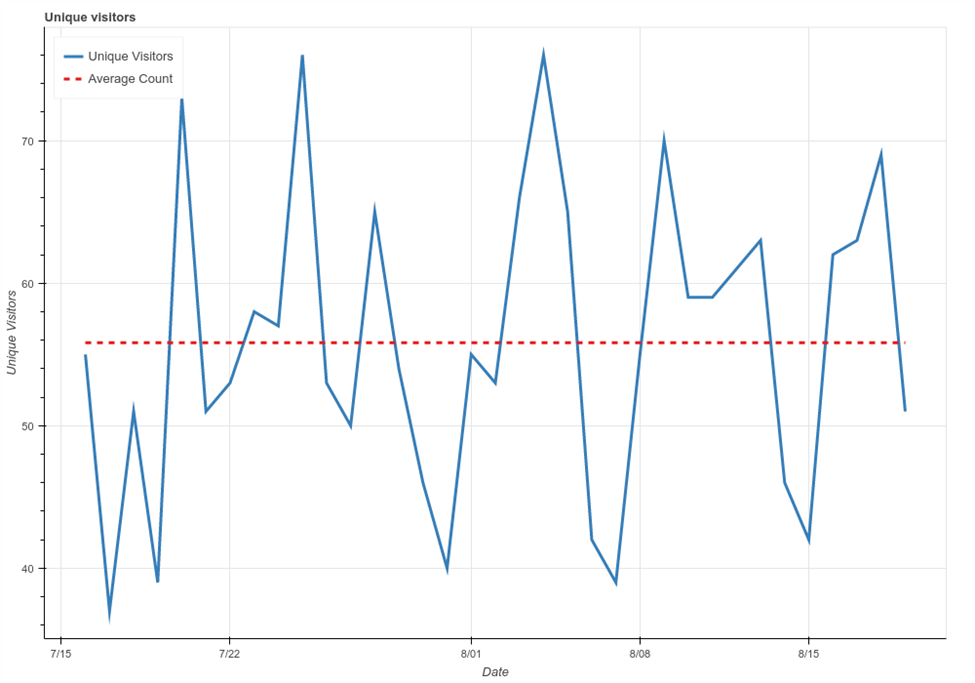


Figure 4 Daily Unique Visitors count

1. **Forecasting daily analytics**

We performed three different forecasting technique on for most visited page, daily 404 response code, daily chapter-1.pdf downloads and unique visitors. We used simple moving average, exponential smoothing and moving window liner regression.

To evaluate the performance of each daily analytics, we used mean absolute error (MAE) and the results are show in table – 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Daily Analytics** | **Moving Average** | **Moving window Linear Regression** | **Exponential Smoothening** |
| most visited page | 10.95 | 13.07 | 10.33 |
| daily 404 response code | 22.23 | 28.38 | 23.43 |
| daily chapter-1.pdf downloads | 13.24 | 14.39 | 15.53 |
| unique visitors | 9.33 | 12.72 | 10.66 |

All the plots of original and different forecasting techniques are shown from figure - 5 to figure – 8.

1. **Conclusion:**

In this report we analyzed most visited pages, daily 404 response code, daily chapter-1.pdf downloads and unique visitors and performed moving average, moving window linear regression and exponential smoothing forecasting technique on the analyzed data. We also given the reason why the number of request for robots.txt is happening and relation of this with 404 code responses.

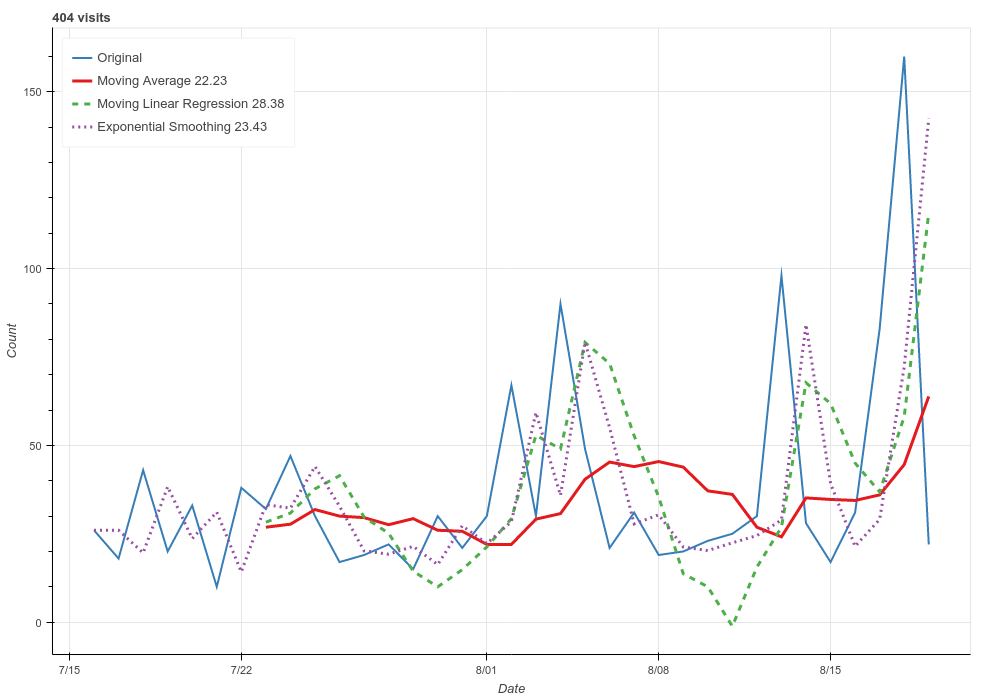


Figure 5 404 code response forecasting

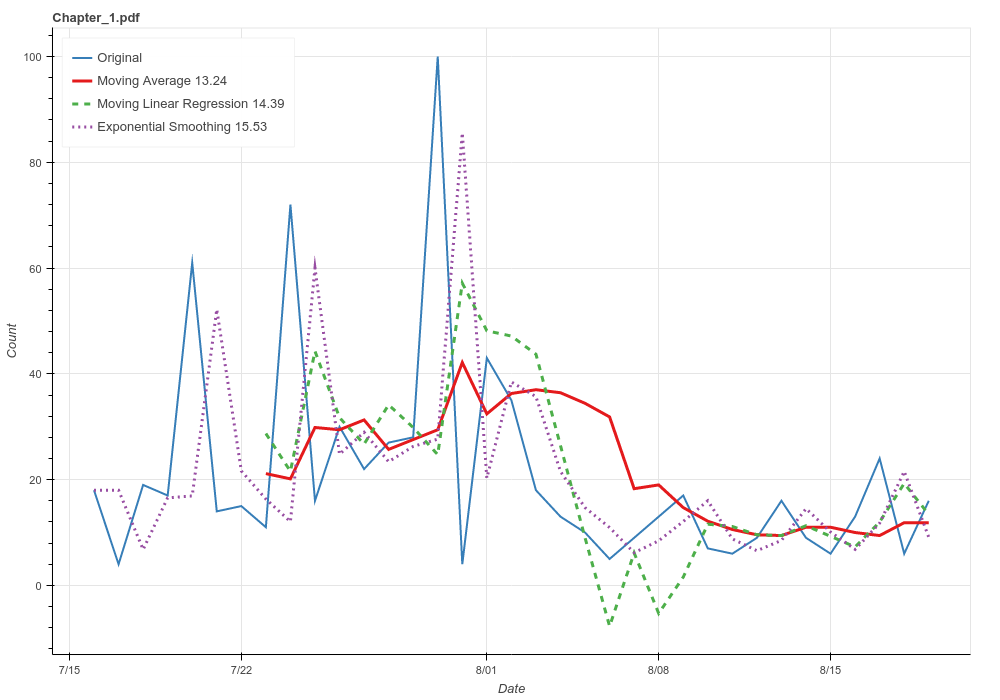


Figure 6 Chapter\_1.pdf forecasting

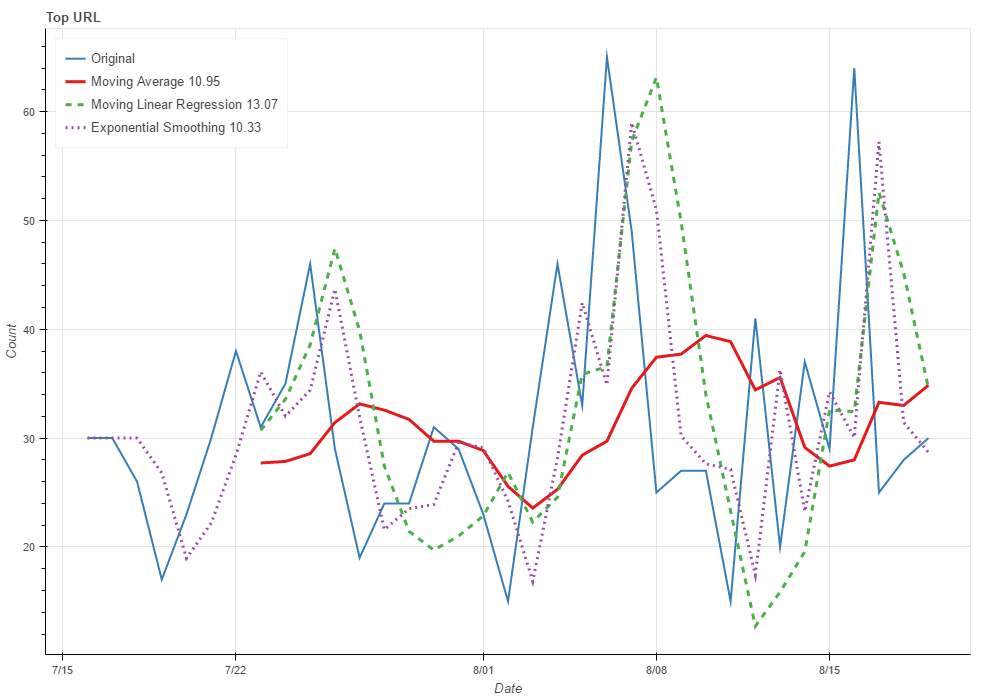


Figure 7 Top-URL forecasting

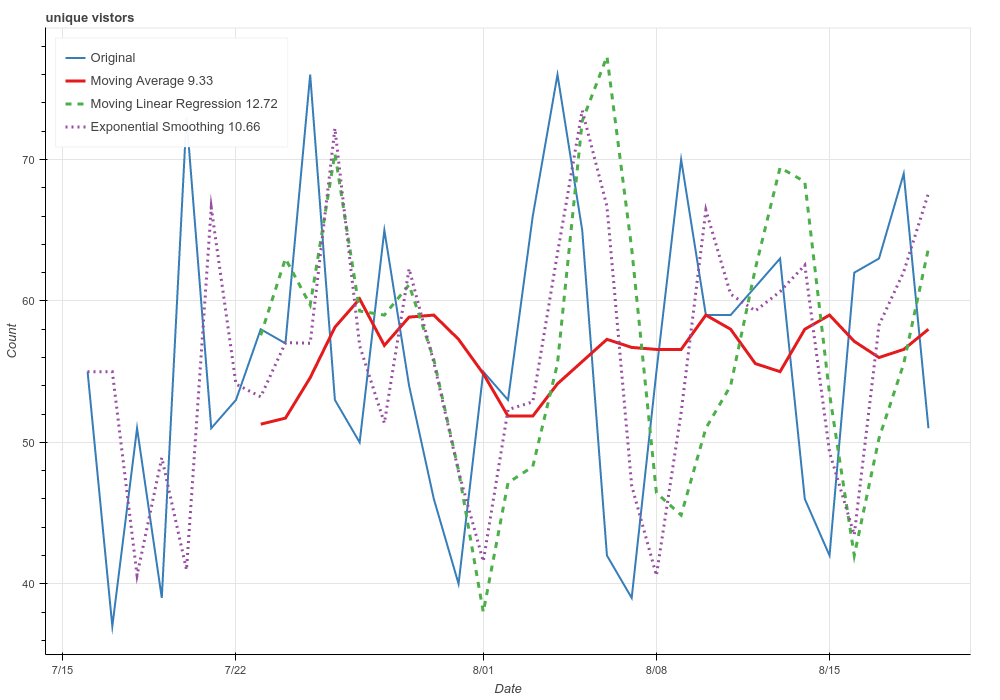


Figure 8 Unique Visitors forecasting