



A Classic Search for Dominant Factors in Web Pages using Search Engines and Validation through Data Analytics using R

J. Sethuraman¹, R. Manikandan¹, K.R. Sekar¹, J. Prassanna² and Ambeshwar Kumar¹

¹School of Computing, SASTRA Deemed University, Thanjavur, India

²School of Computing Sciences and Engineering,
VIT University, Chennai Campus, Tamil Nadu, India

(Corresponding author: R. Manikandan)

(Received 08 April 2019, Revised 29 June 2019 Accepted 16 July 2019)

(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: World Wide Web is plentiful with web pages relevant to different areas. People all over the globe are looking for relevant information via search engines. Capturing and presenting the right information is the utmost job of all search engines. Different search engines use different algorithms to retrieve appropriate web pages. Google uses page rank algorithm to rank pages. Bing has its own proprietary algorithms. Even though algorithms are different these search engines look for certain on page factors in every web page. This paper tries to find the dominant factors which are part of every web page. Majority of search engines always look for these principal factors and the order in which these factors are given weightage differs from one to another. This research work not only captures the governing factors but also tries to validate the selection by means of data analytics. R is used as analytics tool for validation purpose.

Keywords: Search engine optimization, Page Ranking, Google, Bing.

I. INTRODUCTION

Web is affluent with web pages containing relevant information for any topic not only under the sun also over the sun. An automated tool for extracting those pages then to scrap them is needed. A tool for web scraping was used to extract web pages. So, those pages can be used for further analysis. This study focuses on web pages with top keywords from Google trends for the year 2016 those keywords were given to Geturl a self developed tool was used to extract top thirty four pages for an investigational study. And also online tools like "websitetest" and "performanceanalyzer" were used to extract the loading time, performance and overall speed index on the web pages. Those tools expect the web pages as input and it takes a while to come out with the findings. The obtained results were compared with the R tools multiple linear regression. The outputs clearly show that the different characteristics of pages extracted from the thirty sites. The idea is to come out with a new finding on the dominant factors of various sites and how they get searched by the two top search engines. This finding could give an idea on what kind of on page factors can give a better search results to the users.

II. LITERATURE REVIEW

The main objective of this section is to provide a complete review of papers on search engine optimization using various terms. Earlier researchers addressed web site building and design for search optimization techniques [1]. Through Internal coding method and web site overall architecture is analyzed. One of the existing method emphasizes the importance

of SEO to increase the traffic to our sites [2]. Gunjan *et.al* discuss about the ideas of taking a web page to rank better in the search engine results [3]. Wiley provides some tips on search engine optimization for authors to have more citations of their results [4]. Jain discuss about on and off page optimizations [5]. Michael David gives details on recommendations pyramid, SEO in mobiles, impact of social media etc. [6]. The SEO techniques can help search engines rank our sites ahead of competitors [7]. Hussien converse SEO factors like URL, robot meta tags, usage of robot.txt XML sitemap etc., were discussed [8]. Ochoa opines on importance of keywords on title, abstract and keywords and also instructs to build connections with external links etc [9]. Singh and Gupta proposes a new architecture to reduce time and processing power [10]. Some of the important parameters of top web sites were analyzed using R tool as well as the fit of the model with respect to the input data. Prediction deviations between parameters were found for special keywords both Google and Bing search engines gives top ranking, for generic keywords the ranking is low [11].

III. PROPOSED WORK

A self developed tool called Get Url was used and to extract web sites containing pages related to medical images and the partial result is shown in table below The tool was developed using java language and used to extract the given word and key word matching. The frequency of word is calculated based on the location of the keywords within the document. Apart from frequency load time of every page is taken into account. By using multiple online tools were used to find load times and

the average load time is considered. Speed index is another parameter which considers the initial appearance of a web page when it's address is typed in the address bar. Again multiple tools like moz and dynamaper were used and the average of all speed

indexes was calculated. All of the web sites contain the word demonetization as the primary key word. To know the difference between the search results both bing and google search engines were used.

Table 1: Search Results of BING Search Engine Keyword: Demonetization.

S. No	Web Site Name	Freq.	Load time(s)	Speed index
1.	http://www.investopedia.com/terms/d/demonetization.asp	30	4.443	3026
2.	http://www.thefreedictionary.com/demonetization	57	3.301	3168
3.	http://www.dnaindia.com/analysis/column-this-is-a-new-indian-sunrise-2273153	19	6.890	6638
4.	http://demonetization.in/	156	9.599	4054
5.	http://www.dictionary.com/browse/demonetization	52	7.377	6223
6.	http://www.dnaindia.com/topic/demonetization	81	2.230	3011
7.	http://knowledge.wharton.upenn.edu/article/demonetization-india-will-pay-price/	74	5.230	4800
8.	http://www.livemint.com/Opinion/OBjsLy2iZn1Huoyzz1v6WP/Consequences-of-the-demonetization-shock.html	24	3.590	2673
9.	http://www.insightsonindia.com/2016/11/16/big-picture-impact-demonetization/		2.316	1892
10.	http://www.adamimages.com/	82	2.997	1918
11.	https://www.yesbank.in/about-us/important-communication/faqs-demonetization-rbi-directive	74	2.953	2880
12.	http://knowledge.wharton.upenn.edu/article/demonetization-india-will-pay-price/	69	5.647	3588
13.	https://www.yesbank.in/about-us/important-communication/faqs-demonetization-rbi-directive	71	2.808	2756
14.	http://economictimes.indiatimes.com/news/economy/indicators/view-demonetization-modis-biggest-move-is-total-bust/articleshow/60319345.cms	70	2.695	2200
15.	https://taxguru.in/tag/demonetization/	83	3.390	3000
16.	http://www.ndtv.com/topic/demonetization	79	3.611	2930
17.	http://zeenews.india.com/tags/demonetization.html	76	8.580	4789
18.	https://www.project-syndicate.org/commentary/india-demonetization-policy-consequences-by-shashi-tharoor-2016-12	73	4.125	3896
19.	http://www.sify.com/news/10-long-term-goals-of-demonetization-news-columns-glykOnahbbgge.html	81	2.754	2681
20.	https://en.wikipedia.org/wiki/Indian_500_and_1000_rupee_note_demonetisation	69	4.067	3552
21.	http://www.ndtv.com/opinion/demonetization-modis-biggest-move-is-total-bust-mihir-sharma-1744791	74	6.007	6023
22.	http://thediplomat.com/2016/11/the-trouble-with-indias-demonetization-gamble/	71	1.503	1354
23.	https://www.bloomberg.com/view/articles/2017-09-01/demonetization-modi-s-biggest-move-is-total-bust	52	52.639	15743
24.	https://en.wikipedia.org/wiki/2016_Indian_banknote_demonetisation	73	6.584	3052
25.	https://taxguru.in/tag/demonetization/?type=articles	58	5.554	4098
26.	http://timesofindia.indiatimes.com/business/india-business/Unaccounted-deposits-disclosed-to-taxman-face-50-tax-lock-in/articleshow/55620291.cms	59	3.561	3095
27.	https://qz.com/843872/indias-rupee-demonetization-could-spark-a-new-digital-economy-in-the-cash-reliant-country/	64	2.652	2390
28.	http://www.jagranjosh.com/search/demonetization_current-affairs	63	8.443	4534
29.	http://www.moneycontrol.com/news/tags/demonetization.html	70	5.689	8654
30.	https://www.cnbc.com/2016/11/21/india-demonetization-news-expect-short-term-pain-analysts-say-as-growth-expected-to-take-a-hit.html	72	4.568	8653

Table 2: Search Results of Google Search Engine Keyword: Demonetization.

S. No	Web Site Name	Freq.	Load time(s)	Speed index
1.	http://www.investopedia.com/terms/d/demonetization.asp	30	4.443	3026
2.	https://en.wikipedia.org/wiki/2016_Indian_banknote_demonetisation	66	3.301	3168
3.	http://www.dnaindia.com/analysis/column-this-is-a-new-indian-sunrise-2273153	68	6.890	6638
4.	http://demonetization.in/	67	9.599	4054
5.	http://www.dictionary.com/browse/demonetization	52	7.377	6223
6.	http://www.dnaindia.com/topic/demonetization	74	2.230	3011
7.	http://knowledge.wharton.upenn.edu/article/demonetization-india-will-pay-price/	77	5.230	4800
8.	http://www.livemint.com/Opinion/OBjsLy2iZn1Huoyzz1v6WP/Consequences-of-the-demonetization-shock.html	89	3.590	2673
9.	http://www.insightsonindia.com/2016/11/16/big-picture-impact-demonetization/	75	2.316	1892
10.	http://www.adamimages.com/	82	2.997	1918
11.	https://www.yesbank.in/about-us/important-communication/faqs-demonetization-rbi-directive	74	2.953	2880
12.	http://knowledge.wharton.upenn.edu/article/demonetization-india-will-pay-price/	69	5.647	3588
13.	https://www.yesbank.in/about-us/important-communication/faqs-demonetization-rbi-directive	71	2.808	2756
14.	http://economictimes.indiatimes.com/news/economy/indicators/view-demonetization-modis-biggest-move-is-total-bust/articleshow/60319345.cms	70	2.695	2200
15.	https://taxguru.in/tag/demonetization/	83	3.390	3000
16.	http://www.ndtv.com/topic/demonetization	79	3.611	2930
17.	http://zeenews.india.com/tags/demonetization.html	76	8.580	4789
18.	https://www.project-syndicate.org/commentary/india-demonetization-policy-consequences-by-shashi-tharoor-2016-12	73	4.125	3896
19.	http://www.sify.com/news/10-long-term-goals-of-demonetization-news-columns-glykOnahbbgge.html	81	2.754	2681
20.	https://en.wikipedia.org/wiki/Indian_500_and_1000_rupee_note_demonetisation	69	4.067	3552
21.	http://www.ndtv.com/opinion/demonetization-modis-biggest-move-is-total-bust-mihir-sharma-1744791	74	6.007	6023
22.	http://thediplomat.com/2016/11/the-trouble-with-indias-demonetization-gamble/	71	1.503	1354
23.	https://www.bloomberg.com/view/articles/2017-09-01/demonetization-modi-s-biggest-move-is-total-bust	52	52.639	15743
24.	https://en.wikipedia.org/wiki/2016_Indian_banknote_demonetisation	73	6.584	3052
25.	https://taxguru.in/tag/demonetization/?type=articles	65	3.56	2090
26.	http://timesofindia.indiatimes.com/business/india-business/Unaccounted-deposits-disclosed-to-taxman-face-50-tax-lock-in/articleshow/55620291.cms	69	4.56	3090
27.	https://qz.com/843872/indias-rupee-demonetization-could-spark-a-new-digital-economy-in-the-cash-reliant-country/	70	4.6	2590
28.	http://www.jagranjosh.com/search/demonetization_current-affairs	55	3.5	2345
29.	http://www.moneycontrol.com/news/tags/demonetization.html	49	5.6	2454
30.	https://www.cnbc.com/2016/11/21/india-demonetization-news-expect-short-term-pain-analysts-say-as-growth-expected-to-take-a-hit.html	56	5.9	2654

A. Methodology

The tool was used to extract the web pages and the load time and speed index of each page were obtained. The obtained data was stored in two different variables namely google data and Bing data using two different search engines. Then R tool was used and linear regression was applied on both the data. The results obtained are shown in the figure 1 and 2. As R tool is

```
> lmgoogle
Call:
lm(formula = googledata$Load.time.s ~
   googledata$Speed.index +
   googledata$Freq)

Coefficients:
(Intercept)          googledata$Speed.index
googledata$Freq
-1.743309          0.003122          -0.054094

>
plot(googledata$Load.time.s,googledata$Freq,col="blue")
```

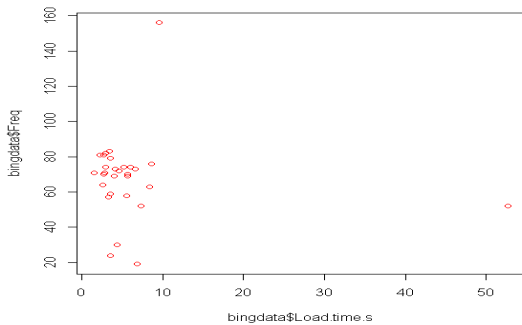


Fig. 1. Bing search engine results.

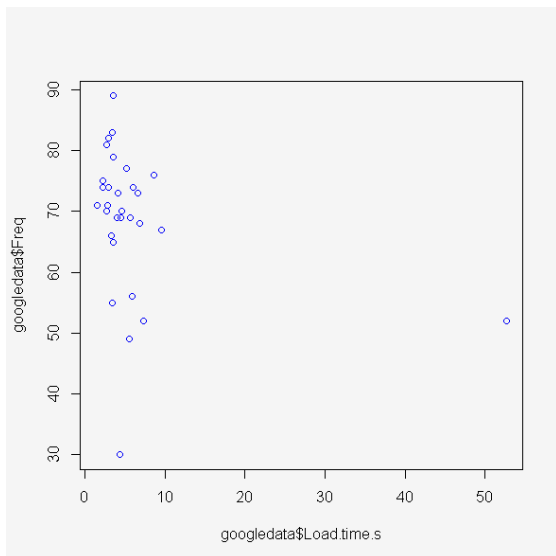


Fig. 2. Google search engine results.

able to apply the algorithm in an effective manner it was chosen. The coefficients and intercept were calculated and also the graph was plotted.

RTool Application on the data set

```
>
lmgoogle=lm(googledata$Load.time.s~googledata$Speed.index+googledata$Freq)

>
lmbing=lm(bingdata$Load.time.s~bingdata$Speed.index+bingdata$Freq)
> lmbing

Call:
lm(formula = bingdata$Load.time.s ~
   bingdata$Speed.index + bingdata$Freq)
Coefficients:
(Intercept)          bingdata$Speed.index
bingdata$Freq
-6.572280          0.002699          0.018089

> plot(bingdata$Load.time.s,bingdata$Freq,col="red")
```

IV. RESULT AND CONCLUSION

Millions and millions of pages are accumulated everyday in the World Wide Web. Indexing the pages is a monotonous job and very difficult in categorizing their segments. Some of the URLs are likely to be available in the first or second page. Search people only see one or two pages for their requirements. This research work brings up the hidden ideology to make our URL to be present and predominantly occupies one of the first two pages. The methodology used in this work revealed and how to bring up the URL to the first few ranking pages. R language is used for the employment of achieving our URL to be present in the first page. The important attribute identified and made our URL to be appeared in the first five ranks. Any search engine will take our URL site to the first page. The remedial measure has been taken in this research work through data analytics and recommender system. The results were obtained and exhibited in the table 1 and table 2. The result clearly shows that Google results give more priority to Speed index and Bing gives more priority to speed frequency of occurrence. Similarity coefficients and index factors were measured through R language and cited above.

Conflict of Interest: Nil

REFERENCES

- [1]. Cui, M., and Hu, S. (2011). Search engine optimization research for website promotion. *International Conference of Information Technology, Computer Engineering and Management Sciences*. Vol. 4, pp. 100-103). IEEE.
- [2]. Beel, J., Gipp, B., and Wilde, E. (2009). Academic Search Engine Optimization (aseo) Optimizing Scholarly Literature for Google Scholar & Co. *Journal of scholarly publishing*, 41(2): 176-190.
- [3]. Gunjan, V.K., Kumari, M., Kumar, A., and Rao, A.A. (2012). Search engine optimization with Google. *International Journal of Computer Science Issues (IJCSI)*, 9(1), 206.

- [4]. Wiley Search engine optimization for authors, Wiley publications 2013.
- [5]. Jain, A. (2013). The Role and Importance of Search Engine and Search Engine Optimization. *International Journal of emerging trends & technology in computer science*, **2**(3): 99-102.
- [6]. Search Engine optimization in 2013, Internetrix Research by Michael David.
- [7]. Omprakash, K.S. (2011). Concept of search engine optimization in web search engine. *International Journal of Advanced Engineering Research and Studies*, **1**(1), 235-237.
- [8]. Hussien, A.S. (2014). Factors Affect Search Engine Optimization, *International Journal of Computer Science and Network Security*, Vol. **14** No.9, 28-33.
- [9]. Ochoa, E.D. (2012). Analysis of the application of selected search engine optimization (SEO) techniques and their effectiveness on Google's search ranking algorithm (Doctoral dissertation, California State University, Northridge).
- [10]. Singh, R., and Gupta, S.K. (2013). Search Engine Optimization-Using Data Mining Approach. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, Vol. **2**, 28-32.
- [11]. Sethuraman, J., Alzubi, J.A., Manikandan, R., Gheisari, M., and Kumar, A. (2019). Eccentric Methodology with Optimization to Unearth Hidden Facts of Search Engine Result Pages. *Recent Patents on Computer Science*, **12**(2), 110-119.

How to cite this article: Sethuraman, J., Manikandan, R., Sekar, K.R., Prassanna, J. and Kumar, A. (2019). A Classic Search for Dominant Factors in Web Pages using Search Engines And Validation through Data Analytics using R. *International Journal on Emerging Technologies*, **10**(2): 117-121.