Web Page Metrics: An Empirical Analysis to Improve the Quality of Web Page

Suman Mann

Abstract— Web Metrics play an important role in measuring the different attributes of a website. It gives actual values of different attributes of website. It can be further used to distinguish between good site design and bad site design. The web page can be evaluated on the basis of different parameter like size of the page, quality of information load time, content available etc. Nowadays website and Internet are emerging media require improvement in their quality for better customer satisfaction. If the website has high page load time or have ambiguous script it results to freeze of web browser due to this user gets irritated and switch to another site. To improve the quality of website and for better understanding we need to measure the website design using the web page metrics. In this paper I gathered the data from Alexa Website and categorize them into good site design and bad site design on the basis of metrics. I have proposed 15 new metrics related to web page design. To achieve our goal we investigate 19 metrics. We present the conclusion of enumerative analysis of web page attributes. The end result of this paper can be used in reckonable studies in designing of web site.

Keywords— Website; Metrics; Web page; Web page quality; Empirical Studies; Web Site Design.

I. INTRODUCTION

Website contents have increased over the time. Earlier web pages used to host images and text contents only but now they include different types of contents like videos, scripts, flash, stylesheet, silverlight etc. Moreover, a website fetches contents from web server where it is hosted and from different third party services like advertising agencies, analytics services etc. Therefore, to render a single web page, it involves fetching of several objects from different servers. Due to this, user faces slow website and generally switches to a different site after experiencing performance issues. Therefore it is very important for web developers to identify the contents of web page which contribute to the bad performance and thus impacts user perceived performance. A web developer has plenty of design recommendations and guidelines for building a usable Web site [12,13,14,15] but still it needs a lot of improvement. We need to maintain the right balance between usability, performance and business interest.

To improve and accelerate the Web site design process there is requirement of new tools and methodologies. New methodology can be develop only if we will measure the correct and efficient web page metrics. Prior to web page metrics researchers had done web measurement by emphasizing on the web graph [3,4], analysis of web traffic[5-10] and analysis of rate of change of content on the web[11].

Even though they all have supported to a better understanding of web measurement and its usage but they do not find the factors which impacts user perceived performance. To analyse any website, metrics play an important role. Since Metrics are vital source of information for decision making there has been a large number of web metrics have been proposed to associate the quality of web page [16].Site reviewer rate web sites on the basis of content, structure, navigation, visual design, functionality, interactivity and overall experience. In order to improve the ranking of website the website developers are required to get the relation between the distinct metrics on the same software for developing the web page. In web site design process we need to characterize the relevant metrics that provide beneficial information otherwise website developer will be confused into so many numbers and the objective of metrics will be lost. Number of web metrics present in the literature is very large and it becomes the tedious process to compute these metrics and get the conclusion and interpret from them. If metrics is properly defined then it can be easier for web developer to make website according to it. We need to understand the metrics for proper designing of websites on which the goodness of website design can depend. In this paper we have introduce 15 new attributes correlated to web page metrics and determine the values of web attributes with the help of tool. This tool is developed in .NET and calculates about 19 web page metrics with significant accuracy.

To meet the above mentioned objective we follow following steps:

- First of all a set of 19 metrics is identified and their values are calculated for top 250 websites of Alexa website.
- Same set of 19 metrics and their values are computed for bottom 250 websites of Alexa website.
- The analysis is drawn to discover the subset of attributes which can relate to goodness of website design.
- These attributes can be used to evaluate the data into good design and bad design.

The objective of this paper is to discover the metrics from available and newly introduced to get the benchmark of good design. Our paper is organized as follows: In Section II of the paper the metrics of web page is tabulated which is used in our paper. Section III describes the methodology used in the research, collection of the data and tool description which we used for evaluating the attributes of the web page. In Section IV we explain the approach that we used to analyze the data and present the result. In Section V conclusion is discussed. Future work is specified in section VI.

Revised Version Manuscript Received on February 15, 2016.
Suman Mann, Maharaja Surajmal Institute of Technology, New Delhi, India.

II. EXPLANATION OF METRICS OF WEB PAGE SELECTED FOR THE STUDY

Even though, many researcher has proposed number of metrics [15,17,18,19] for web page, out of these we identified few metrics and introduce 15 new metrics for our study that are tabulated in table 1.Prior to our work there are 42 metrics already defined for web page and classification of those is given as follows:-

- Metrics for Page composition:-This metrics include Number of words, Total number of Body Text words, Total Words in page title, Total number of links etc.
- Metrics for Page formatting: - This metrics comprise of Font size, Font style, Screen coverage etc.
- Overall page assessment or quality metrics: - This metrics include Quality of Information, Quality of Image, Quality of Link etc.

The metrics for overall page quality cannot be easily computed as they require human involvement. Therefore in our study, we only take Page formatting metrics and page composition metrics which can be simply calculated.

Table I. Web Page Metrics

Download Time(s)	Time to Download the page in sec.
Total Content Load(kB)	Total Size of content loaded for a page
Number of Requests	Total number of request for loading a page
Size of Images(kB)	Total Kilobytes of images on a page
Size of Scripts(kB)	Total Kilobytes of scripts running on page
Size of Css(kB)	Total Kilobytes of style sheet used in a page
Size of Flash(kB)	Total Kilobytes of Flash content in a page
Size of HTML(kB)	Total Kilobytes of HTML content for page
Percentage of jpeg images	Total percentage of jpeg images compared to total images of a page
Percentage of png images	Total percentage of png images compared to total images of a page
Percentage of gif images	Total percentage of gif images compared to total images of a page
Percentage of other images	Total percentage of other types of images compared to total images of a page
Percentage of https	Total percentage of https content on a page
Percentage of http	Total percentage of http content on a page
Number of Words	Total words on a page
Total Links	Total Links available on a page
Number of Internal Links	Total Internal links embedded on a page
Number of External Links	Total External links on a page
Number of Broken Links	Total Broken links on a page

The explanation of the parameters used in our study is as follows:-

1. Download Time

We have taken the total time in seconds to load a page. This attribute is calculated by observing total time to download a page including all contents.

2. Total Content Load(kB)

This metric specifies the total size of content loaded for displaying a web page.

3. Number of Request

This metrics counts the total number of request by a page to completely load. In this, we calculate the number of request to third party content also.

4. Size of Images

This metrics evaluate the total size of images present on the page.

5. Size of Scripts(kB)

This metrics calculate the size of scripts present on the page. Scripts are like javascript etc.

6. Size of Css(kB)

This metrics evaluate the size of cascading style sheet that is used in the web page.

7. Size of Flash(kB)

This metrics calculates the size of flash content embedded on the page.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

8. Size of HTML(kB)
This metrics evaluate the size of html code written for a page.
9. Percentage of jpeg images
This metrics calculate the total percentage of jpeg images present as compared to total images in the page.
10. Percentage of png images
This metrics evaluate the total percentage of png images present as compared to total images in the page.
11. Percentage of gif images
This metrics calculate the total percentage of gif images present as compared to total images in the page.
12. Percentage of other images
This metrics evaluate the total percentage of other type of images present as compared to total images in the page.
13. Percentage of https
This metrics calculate the https content present in the page.
14. Percentage of http
This metrics evaluate the http content present in the page.
15. Number of Words
This metrics calculate the number of words present in the page.
16.Total Links
This metrics counts the total number of links on a web page and can be calculated by counting the number of links present on the web page.
17.Number of Internal Links
This metrics evaluate the total number of links to the other page of website.
18. Number of External Links
This metrics evaluate the total number of links to the other external website.
19.Number of Broken Links
This metric calculate the total number of broken link present on the page.

III. RESEARCH APPROACH

This paper evaluate enumerative metrics of web page for example download time, total content load, number of requests, size of different types of images, percentage of different types of images on page, number of links, number of words etc from the web pages that was evaluated for 500 websites.
Since 2014, Alexa provides traffic data, global rankings and other information on 30 million websites and its website is visited by over 8.8 million people monthly.
Alexa Internet was founded in 1996 by American web entrepreneurs Brewster Kahle and Bruce Gilliat. The company's name was chosen in homage to the Library of Alexandria, drawing a parallel between the largest repository of knowledge in the ancient world and the potential of the Internet to become a similar store of knowledge.
Alexa Internet, Inc. is a California-based subsidiary company of Amazon.com which provides commercial web traffic data. Founded as an independent company in 1996, Alexa was acquired by Amazon in 1999. Its toolbar collects data on browsing behavior and transmits it to the Alexa website, where it is stored and analyzed, forming the basis for the company's web traffic reporting.

A. Gathering of Data

All the web sites are taken from Alexa. We have collected top 250 web sites and bottom 250 websites. We have taken the home page only for evaluating distinct web pages. Website can be categorized into three levels in the as first level, second level and third level pages. First level page is the home page. The second level comprise of pages that are accessible directly from first level that is home page and the third level pages are accessible from second level but not from the home page. We have taken only first level page in this paper. The Data gathering procedure is explained in the figure 1 of block diagram.

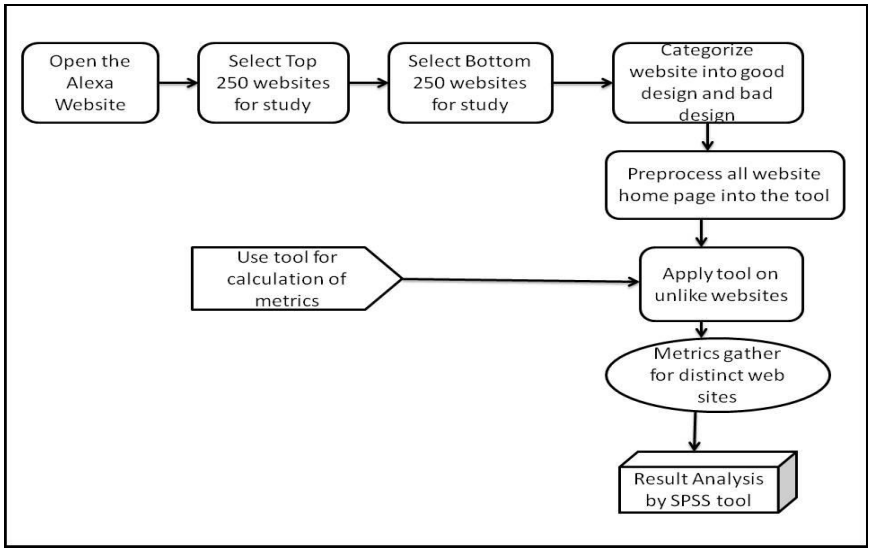

Figure 1: Block Diagram of Data Gathering Procedure

B. Tool description

The study of web page metrics is automated by the tool we develop for evaluating 19 web page attributes. We use .NET framework using C# language for this purpose. .NET framework is one of powerful, easy to use and fundamental tools in a Web site developer's toolbox. The .NET framework was introduced by Microsoft in June 2000 with a vision for embracing the internet. It is language and platform independent and has

tie-up with the Windows operating System. We can use various languages like visual basic .NET, Visual C++ .NET, C#, Iron Python and more, and executes program irrespective of the language used with the Common language Runtime (CLR). .NET programs compiled to Microsoft Intermediate Language (MSIL), which is then translated into machine code and executed. The tool we mentioned above can evaluate different web attributes. We can select some of the above list or select all the attributes. The tool interface is shown in figure 2.

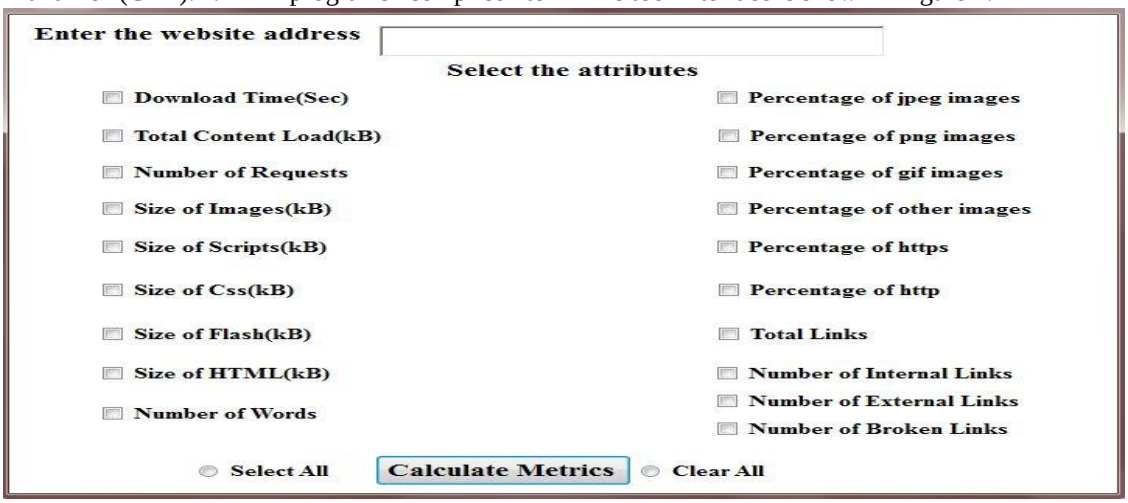

Figure 2: Interface Diagram of Tool

IV. RESULT ANALYSIS

In this portion we explain the procedure used to explore the metrics data calculated for 500 web sites. We used statistical techniques to explain the nature of the 500 website data of the year 2004 Alexa. We have also used Logistic Regression to discriminate good page design from bad page design. This method is suitable where we have only one dependent variable and many independent variables. In our study, we have only one dependent variable good design /bad design and whole web metrics are the independent variables. Logistic Regression is widely used procedure to analyze the data. It is used to predict dependent variable from a set of independent variables. The dependent variable in our study is good design and bad design and the web metrics are independent variables. In Logistic Regression forward selection and backward elimination can be used. Stepwise entry variable examines the variable that is selected one at a time for entry at each step. This is a forward stepwise procedure. The backward elimination method includes all independent variables in the model. Variables are deleted one at a time from the model until stopping criteria are fulfilled. We have used forward selection method to explore top 250 websites data and backward elimination method for 250 website data. Based on the above calculations the values calculated can be tabulated as follows:

Table II-Top 250 Websites

Metrics	Good design			
	Minimum	Maximum	Mean	Std. Deviation
Download Time(s)	0	24	**5.22**	5.413
Total Content Load(kB)	96	4623	**1214.03**	937.957
Number of Requests	15	246	**92.59**	71.848
Size of Images(kB)	24	3932	**768.08**	815.102
Size of Scripts(kB)	0	936	**231.36**	169.915
Size of Css(kB)	0	343	**39.05**	60.138
Size of Flash(kB)	0	246	**25.05**	53.969
Size of HTML(kB)	3	379	**66.69**	75.545
Percentage of jpeg images	0	87	**29.41**	23.322
Percentage of png images	2	94	**41.54**	24.043
Percentage of gif images	0	58	**24.13**	16.112
Percentage of other images	0	19	**4.72**	5.236
Percentage of https	0	100	**32.21**	43.635
Percentage of http	0	100	**67.79**	43.635
Number of Words	0	1972	**140.13**	410.286
Total Links	1	2107	**240.31**	410.178
Number of Internal Links	0	827	**82.49**	161.729
Number of External Links	0	1853	**147.05**	340.282
Number of Broken Links	0	34	**5.10**	8.807

Table III-Bottom 250 Websites

Metrics	Bad Design			
	Minimum	Maximum	Mean	Std. Deviation
Download Time(s)	1	26	6.06	4.845
Total Content Load(kB)	30	13030	1703.79	2354.114
Number of Requests	6	844	107.42	133.438
Size of Images(kB)	11	7168	810.43	1253.845
Size of Scripts(kB)	0	1256	287.25	219.394
Size of Css(kB)	0	409	54.85	75.621
Size of Flash(kB)	0	2672	96.85	377.658
Size of HTML(kB)	1	255	41.76	43.558
Percentage of jpeg images	0	89	33.43	27.666
Percentage of png images	0	86	39.97	26.262
Percentage of gif images	0	63	20.64	18.002
Percentage of other images	0	588	14.15	69.058
Percentage of https	0	99	27.64	40.470
Percentage of http	1	100	72.36	40.470
Number of Words	0	2868	221.60	550.355
Total Links	1	983	206.13	245.757
Number of Internal Links	0	981	117.82	196.901
Number of External Links	0	849	88.44	169.982
Number of Broken Links	0	294	15.69	46.057

V. CONCLUSION

The goal of our research is to capture good quality of web sites. Nowadays E-business is emerging and websites are not mere a medium for communication they become a product for providing services. Therefore imparting quality, security and reliability to web sites are very important. We empirically validate the relationship of web metrics and quality of websites using logistic regression technique the results are based on Alexa data for 2014. In this paper we present the attributes which, if have higher value can lead to a bad design. From the above attributes we also find profile of good pages. The type of metrics explored here are only one piece of the web site design puzzle; this work is part of a larger project whose goal are to develop techniques to empirically investigate all aspect of web site design and to develop tools to help designers of the web site to improve the quality of the web page.

The empirical data obtained by using the program reveals the following things for good website design.

1. Download Time(s) should be low.
2. Total Content Load (kB) should be low.
3. Number of Requests should be low.
4. Size of Images (kB) should be low.
5. Size of Scripts (kB) should be low.
6. Size of Css (kB) should be low.
7. Size of Flash (kB) should be low.
8. Size of HTML (kB) should be High.
9. Percentage of jpeg images should be low.
10. Percentage of png images should be High.
11. Percentage of gif images should be High.
12. Percentage of other images should be low.
13. Percentage of https should be High.
14. Percentage of http should be low.
15. Number of Words should be low.
16. Total Links should be High.
17. Number of Internal Links should be low.
18. Number of External Links should be High,
19. Number of Broken Links should be low.

VI. FUTURE WORK

In future I will replicate this work on the larger set of data and will explore new tools and methodology in all dimensions. Using that work of web site engineers will simplify and quality of the web sites will improve. In future I will explore and take first level and second level web pages because the characteristics of home page is different from other levels of the page .In future I will propose guidelines for making effective web sites which can be easily downloaded and have good scan ability.

REFERENCES

1. Alexa. http://www.alexa.com/.
2. HTTP archive. http://httparchive.org/.
3. A. Broder et al.," Graph structure in the web. Computer Networks", 33(1), June 2000.
4. J. M. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan et al.," The web as a graph: Measurements, models and methods",In Proc. COCOON, 1999.
5. B. Krishnamurthy, C. E. Willis et al.," On the use and performance of content distribution network"In Proc.IMW, 2001
6. S. Singh et al. "Active measurement system for high-fidelity characterization of modern cloud applications" In Proc. USENIX Conference on Web Applications, 2010.
7. F. Schneider, S. Agarwal, T. Alpcan et al.,"The new Web: Characterizing AJAX traffic" In Proc. PAM, 2008.
8. A. Nazir, S. Raza, D. Gupta,and B. Krishnamurthy," Network level footprints of Facebook applications" In Proc. IMC, 2009.
9. F. Schneider, A. Feldmann, B. Krishnamurthy et al,"Understanding online social network usage from a network perspective" In Proc. IMC, 2009.
10. P. Gill, M. Arlitt, N. Carlsson and C. Williamson.,"Characterizing Organizational Use of Web-based Services: Methodology,Challenges,Observations, and Insights" ACM TWEB, 2011.
11. D. Fetterly and J. Wiener, " A large scale study of the evolution of web pages" In Proc.WWW, 2003.

12. Vincent Flanders and Michael Willis,"Web Pages That Suck: Learn Good Design by Looking at Bad Design" SYBEX, San Francisco, 1998.
13. Jakob Nielsen, "The alertbox: Current issues in web usability", http://www.useit.com/alertbox.
14. Jakob Nielsen, " User interface directions for the Web," Communications of the ACM, 42(1):65–72, January 1999.
15. Jakob Nielsen, "Designing Web Usability: The Practice of Simplicity", New Riders Publishing,Indianapolis, IN, 2000.
16. Karen A. Shriver, "Dynamics in Document Design", Wiley Computer Publishing, John Wiley & Sons, Inc., New York, 1997.
17. Lincoln D. Stein, "The rating game", http://stein.cshl.org/lstein/rater/,1997.
18. George W. Furans, "Effective view navigation", in proceedings of ACM CHI 97 conference on human factors in computing systems, volume 1 of PAPERS: information structures, pp. 367-374, 1997.
19. Kevin Larson and Mary Czerwinski., "Web page design: Implications of memory, structure and scent for information retrieval", In proceedings of ACM CHI 98 Conference on human Factors in Computing Systems,volume 1 of Web Page Design , pp. 25-32, 1998.