International Journal of Computer Science and Engineering Open Access
Survey Paper Volume-2, Issue-2 E-ISSN: 2347-2693

A Study on Website Quality Evaluation based on Sitemap

Chandran M[1], Ramani A.V[2]

[1*]*MC Department, SRMV College of Arts and Science, India, onchandran@gmail.com,*
[2] *Computer Science Department SRMV College of Arts and Science, India, avvramani@yahoo.com*

www.ijcaonline.org

Received: 03 Feb 2014 Revised: 17 Feb 2014 Accepted: 26 Feb 2014 Published: 28 Feb 2014

***Abstract*—** Website quality evaluation can be made based on creating site map for the WebPages for a single website which works properly. A website is taken for the analysis where we check every link under that website are checker and split it according to status code. By analyzing every status code from all those webpage links, we are ignoring every other link except the page contains status code 200. Then we are developing the sitemap for those links which is working perfectly.

***Keywords*—** *Sitemap, Website, Search Engine Optimization*, SMGA

I. INTRODUCTION

Website is something entirely new in the world of software quality, within minutes of going live. The World Wide Web has made the spread of information and ideas easy and accessible to millions. It's the place where everyone has an opportunity to be heard—that is, if you can be found amidst the multitude of other Web sites out there. Every WebPages has their own characteristics and this characteristic has drawbacks and benefits.[1]

There are many dimensions of quality, and each measure will pertain to a particular website in varying degrees. Here are some of them: time, a credible site should be updated frequently. The information about latest update also should be included on the homepage. However, if the information has not been updated currently, the visitor could easily know that perhaps the site manager does really bother to update the site. Second is structural, all of the parts of the website hold together and all links inside and outside the website should work well. Broken links on the webpage also are another factor that always downgrades the quality of website. Each page usually has references or links or connections to other pages. These may be internal or external web site. A user expects each link to be valid, meaning that it leads successfully to the intended page or other resource. In a 2003 experiment, discovered that about one link out of every 200 disappeared each week from the Internet [1]. The third factor is content; number of the links or link popularity is

One of the off page factors that search engines are looking to determine the value of the webpage. Most of search engine will need a website to have at least two links pointing to their site before they will place it to their index, and the idea of this link popularity is that to increase the link popularity of a website, this website must have large amount of high quality content. Number of links to website improves access growth and helps to generate traffic [2].

$$PR(A) = (1-d) + d(PR(t1)/C(t1) + ... + R(tn)/C(tn))$$

PR = page rank

t1 – tn = are pages linking to page A

C = is the number of outbound links that a page as

D = is a damping factor, usually set to 0.85.

Search engine such Google make a citation analysis to rank hits, then a website which has a any links to it will have a higher ranking compare than a website with a few links. This indicator can be used to measure the quality of web site. Fourth is response time and latency, a website server should respond to a browser request within certain parameters, it is found that extraneous content exists on the majority of popular pages, and that blocking this content buys a 25-30% reduction in objects downloaded and bytes, with a 33% decrease in page latency. Popular sites averaged 52 objects per page, 8.1 of which were ads, served from 5.7 servers [3], and object overhead now dominates the latency of most web pages [4]. Following the

The first step would be to be sure your sitemap is up to date to begin with - and has all the URLs you want. The main thing is none of them should 404 and then beyond that, yes, they should return 200's. Unless you're dealing with a gigantic site which might be hard to maintain, in theory there shouldn't be errors in sitemaps if you have the correct URLs in there. Getting sitemaps right on a large site made a huge difference to the crawl rate and a huge indexation to follow [3]. With growth of web-site content it's getting harder and harder to manage relations between individual

Corresponding Author: *Chandran M, onchandran@gmail.com*

WebPages and keep track of hyperlinks within a site. Unfortunately there are no perfect web-site integrity tools or services that can enforce proper relationship between pages, keep track of moving content, webpage renames etc, and update corresponding URLs automatically. Modern content management systems and blog software may aggravate the problem even more - by Replicating the same dead web links across numerous web-pages which they generate dynamically, so people can be getting 404 errors much more frequently.[4]

Sitemap

Sitemaps, as the name implies, are just a map of your site - i.e. on one single page you show the structure of your site, its sections, the links between them, etc. Sitemaps make navigating your site easier and having an updated sitemap on your site is good both for your users and for search engines.[3].

Important sitemap errors that could affect our rankings

The first step would be to be sure your sitemap is up to date to begin with - and has all the URLs you want (and not any you don't want). The main thing is none of them should 404 and then beyond that, yes, they should return 200's. Unless you're dealing with a gigantic site which might be hard to maintain, in theory there shouldn't be errors in sitemaps if you have the correct URLs in there. Getting sitemaps right on a large site made a huge difference to the crawl rate and a huge indexation to follow.

II. PROBLEM DEFINITION

Every webpage design has their own characteristics and this characteristic has drawbacks and benefits. There is a mechanism for measuring the effects of the webpage component toward the performance and quality of website. This mechanism will measure size, component, and time needed by the client for downloading a website. The main factor that will influences this download time are page size (bytes), number and types of component, number of server from the accessed web. Research conducted by IBM can be used as a standard for performance measurement of quality [7]. Standard international download time for this performance can be used as a reference to categorize the tested webpage. After we have done with data, and then continued by testing of data.

Table1. Standard of the website performance

Tested Factor	Quality Standard
Average server response time	< 0.5 second
Number of component per page	< 20objects
Webpage loading time	< 30 second
Webpage size in byte	< 64 k

Four reasons to keep site map

A site map is literally a map of your Web site. It is a tool that allows visitors to easily get around your site. Having a well constructed site map is not only important to create a positive experience for your potential customers, but is an important aspect of search engine optimization. Below are 4 functions of a site map.

Navigation

A site map provides an organized list of links to all the pages on your Web site. If visitors get lost while browsing your site, they can always refer to your site map to see where they are and get where they would like to go. Site maps allow your visitors to navigate your Web site with ease.

Theme

When visitors access your site map, they will learn a lot about your Web site within a very short period of time. A well constructed site map will allow visitors to easily and efficiently grasp your site.

Search Engine Optimization (SEO)

Since a site map is a single page that contains links to every page on your Web site, it is a very effective way to help search engine spiders crawl through your site with ease. Since search engines rely on links to find the main pages of your site, a site map is a great way to get every page on your site indexed by the search engines. The more pages you have indexed by the search engines, the more potential you will have to reach a greater number of prospective clients. The World Wide Web has made the spread of information and ideas easy and accessible to millions. It's the place where everyone has an opportunity to be heard—that is, if you can be found amidst the multitude of other Web sites out there.

Search Engine Optimization (SEO) is the process of making your Web site accessible to people using search engines to find services you provide. Search engines (such as Google, Yahoo, and Bing) operate by providing users with a list of relevant search results based on keywords users enter. This allows people who don't know your Web site address to find your site through keyword searches [1].

Some basic features of Web sites that search engine spiders look for are: Keyword usage, Keyword placement, Compelling content, HTML title tags, meta-descriptions and Keyword tags, External and internal links, Site updates, Site map, Web design, Functionality. Effective keyword usage is not simply based on repeating a keyword or phrase over and over on your Web site.

Organization

A site map enables you to easily assess the structure of your site to see where your site is strong and where it is weak. Whenever you need to add new content or new

sections to your Web site, you will be able to take the existing hierarchy into consideration by glancing at your site map.[1]

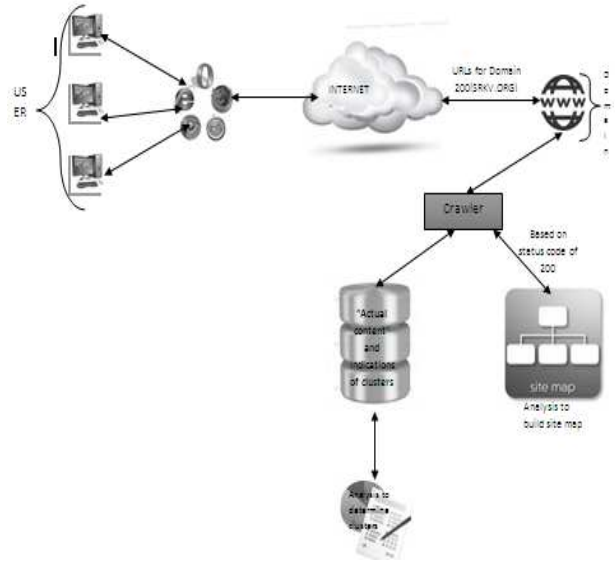
Sitemap files have a limit of 50,000 URLs and 10 megabytes per sitemap. Sitemaps can be compressed using gzip, reducing bandwidth consumption. Multiple sitemap files are supported, with a Sitemap index file serving as an entry point. Sitemap index files may not list more than 50,000 Sitemaps and must be no larger than 10MiB (10,485,760 bytes) and can be compressed. You can have more than one Sitemap index file [2]

III. METHODOLOGIES

This research stages will start with problem identification followed by research procedure and sample of data explanation.

Nature of invalid hyperlinks

With growth of web-site content it's getting harder and harder to manage relations between individual WebPages and keep track of hyperlinks within a site. Unfortunately there are no perfect web-site integrity tools or services that can enforce proper relationship between pages, keep track of moving content, webpage renames etc, and update corresponding URLs automatically. With time this causes some hyperlinks become obsolete, stale, dangling, and simply - dead because they don't lead to valid pages anymore, and web-users are going to get 404 response codes or other unsuccessful HTTP responses each time when they try to access the web-pages. Modern content management systems and blog software may aggravate the problem even more - by replicating the same dead weblinks across numerous web-pages which they generate dynamically, so people can be getting 404 errors much more frequently.



Important of online link checker

Due to lack of adequate problem detection tools (aka URL validators, web spiders, HTML crawlers, website's health analyzers etc) it's really very hard to identify what exact local and outbound hyperlinks became dead, and it's even harder to fix those because in order to do so you need to know precise location of the broken linking tag in the HTML code: without that you will need to scan through thousands source lines to find exact HREF (or other linking sub-tag) that causes the problem.

Sample Data

In order to get data for this research, we examined Ramakrishna mission portals were not randomly selected, but a careful process was undertaken. Rather than selecting any generic [5]

At the beginning of the process we are giving the website link. As we can see that the status of that website, whether it presents or not. By the analysis of this functionality we can able to get the status code of the website link. As shown in the figure, the domain name, ip address and server name with status code would be displayed. If the website status code is 200 then the website link that we gave is completely ok. If the website link we gave is broken or deleted than it will display the 404 status code error.

Constructing Tree Structure by Applying Site Mapping Generation Algorithm (SMGA)

MAPGEN(D_i)

{

GenRoot(m_1,....m_n);

// Getting root node for menus

For i $\rightarrow$0 to n

For j $\rightarrow$0 to f

s[j]=GetChild(m_i,j);

// For getting child node

End For

If s[i]==NULL

AddNode(m_i,NULL);

//No child node for root

Else

AddNode(s[i], m_i);

// adding child to root node

End If

End For

For all m, s∈ Domain

}

IV. Result and Discussion

In Table-2, we are giving a website link (http://www.srkv.org) to test whether that link present or not. After receiving that the status of that link as 200, we are examining whether that link has site map or not. If we got to know that the website does not have the site map, we are moving to the next step of process.

Table-2:

ID	PageName	Status Code	DomainName	IpAddress	ServerName
1	http://www.djmaza.com	200	www.djmaza.com	50.7.44.146	nginx admin
3	http://www.srmvcas.org	200	www.srmvcas.org	174.120.8.222	Apache
4	http://srkv.org	200	www.srkv.org	108.162.196.114	cloudflare-nginx
5	http://www.b-u.ac.in	200	www.b-u.ac.in	5.10.76.145	Microsoft-IIS/6.0

From Table-3, we are exploring how many links totally that website contains. With the help of that data we are processing every single link that we got to receive the status code of that link. By categorizing that we are splitting them into number of collection sorted by status code. We are developing the site map for the link which has the status code 200. We are ignoring the rest of the links from that website.

Table3. Dynamic website (www.srkv.org) List of errors with status code.

S.NO	Status Code	Number of Errors
1	200	84
2	404	104
3	410	4
4	522	1

Table-4 shows the common status code that occurs often with description and comment.

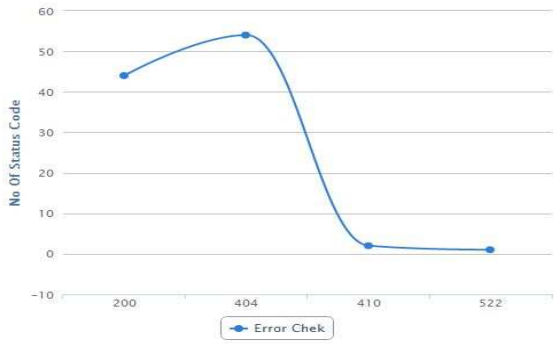
When the received links which has the status 200, we could confirm that the link of that website link is working fine. When the received code is 404, the requested page or the URL is not available or unknown location to the server. When the received status code is 522, the requested web server is currently down or unavailable due to traffic.

Table-4:

S.NO	Status Code	Discription	Comment
1	200	OK	Action completed successfully
2	404	Not Found	The requested file was not found.
3	410	Gone	The requested page is no longer available
4	522	web server	The request was able to connect to your web server

Figure-1 represents in form of chart which the data that collected from the Table-2. From the chart we can understand that the status code of the website link data has drawn where the 404 status code occurs often than others.

Figure-1

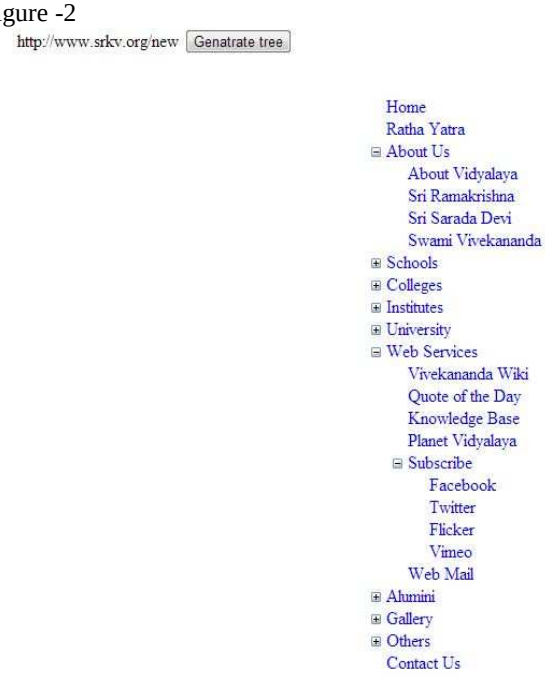


First step for creating site map we need a site to analyse the WebPages under that site. For that we are taking a link (www.srkv.org) for creating site map. After reading every page under that link, we can get a table of content which has a series of links with the status code. We found that the total link contains under that website is 193. By categorizing those pages according to the status code of the every link. The total link that contains 200 Status code is 84. The total link that contains the 404 status code is 104. The total link that contains 410 Status code is 4. The total link that contains Status code 522 is 1. By ignoring all the links that contains status code except 200. We can only create site map for the link which contains the 200 status code.

Development of the sitemap Generator

The sitemap which shown in the Figure-2 had generated with the help of above algorithm. The algorithm defines the process to show the result by means of root node and child node. If the link is Title it adds to the root node else it add that sub title link as child node.

Figure -2



V. CONCLUSION

A required link of the website to check all the links under that which has the status code 200 and ignoring the remaining error links. The page which works completely has been taken for creating sitemap based on SMGA algorithms. Search Engine optimization Looks for sitemap in every website for the ranking system in every query search. We are developing the sitemap for the website which already do not have the sitemap within. When the SEO found the sitemap in a website then it would increase the ranking

REFERENCES

[1]. Frank McCown, M.N., and Johan Bollen, The Availability and Persistence of Web References in D-Lib Magazine, in the 5th International Web Archiving Workshop and Digital Preservation (IWAW'05). **2005**.

[2]. Larry Page, R.M., Sergey Brin,Terry Winograd. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Stanford.

[3]. Krishnamurthy, B.a.C.W. Cat and Mouse: content Delivery Tradeoffs in Web Access. in WWW 2006. Edinburgh, Scotland.

[4]. Yuan, J., Chi, C.H., and Q. Sun. A More recise Model for Web Retrieval. in WWW 2005. **2005**. Chiba, Japan.

[5]. Team, I.W.S., Design for Performance: Analysis of Download Times for Page Elements Suggests Ways to Optimize. **2001**.

[6]. Information on "Helping Spiders Crawl through your Web Site "available at, http://sonicseo.com/helping-spiders/last accessed at 18 September, **2013**.

[7]. Information on "sitemaps" http://en.wikipedia.org/wiki /Sitemaps#Sitemap_index/,last modified on 15 September **2013**,

[8]. Information on" Free Broken Link Checker /OnlineURLValidator"http:/brokenlinkcheck.com/, last accessed at 18 September, **2013**

[9]. Handaru Jati and Dhanapal Durai Dominic "Quality Evaluation of E-Government Website using Web Diagnostic Tools:Asian Case",2009 International Conference on information management and Engineering , **2009** IEEE.

AUTHORS PROFILE

Author is an Assistant Professor in Computer Applications Department at SRMV College of Arts and Science, Coimbatore, Tamil Nadu. He has many Publications in national & international journal and Conferences. Currently he is doing research in the area of Software Engineering