ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Finding Similarity of Web Sites Based On Analogous Feature Set

Sethuraman J*, and Vaithiyanathan V.

SASTRA University, Thanjavur, Tamil Nadu, India.

ABSTRACT

As the number of Internet users is growing exponentially, the number of web sites containing web pages are also increasing rapidly. But this increase has made the job of users and search engines very difficult in choosing the appropriate content. This article is an attempt to find web pages with similar characteristics and finds what characteristics make them similar, and find the correlation between the selected features. With the findings it is easier for web developers to incorporate the necessary changes in the web site to promote them on top of search engine results.

Keywords: Search Engines, Web pages, optimized contents, Search engine result pages (SERPS), organic search, GS(globalscore), ACCSAccessibility), DSG(desgin), TXT(Text),MM(Multimedia),NW(Network)

**Corresponding author*

INTRODUCTION

Internet usage has become an indispensable part of everybody and everyone has started using it for everything right from general search to specific fulfillments like online transactions. General search can be with respect to some online remedies to small health issues, educational topic searches, public searches for a specific location of a place etc. Specific searches can be with respect to loan, mortgage, finance etc., all these searches now days result in millions of fetches through search engine results where many a times web pages with reliable contents are ignored and web pages with fake contents are brought to top of the lists. Search engines do their best to bring out the most reliable contents to the top. But it is the Search engine optimization researcher's responsibility to suggest the best techniques to be followed while developing the web pages keeping in mind the content of the web pages so that those pages are resulting in organic search results.

Existing work

The following survey lists out the various factors identified in different research articles and white papers on SEO techniques. Elsevier's Getnoticed [1] Shows the ways and techniques to be followed by research scholars for improving their articles rankings with search engines. Seema Rani et al. [2] review the importance of online SEO techniques for improving the importance of web site ranking. Ceonex, Inc [3] suggests getting links to web sites, appropriate keyword selection by doing keyword research, simple and concise url ,submitting url to search engines, relevancy of content, avoiding search engine stumbling blocks, submitting key pages, keeping the contents and listings fresh, using site map and every page of a web site should be within three clicks from the home page. Google Scholar's Ranking Algorithm [4] stresses on the importance of keywords appearance in title, article's citation count, age of an article, search term occurrence in article's title, in author or publication name, occurrence, frequency of search term in article's full text , wasfa kanwal [5] says web sites with unique, quality and up-to-date contents with more back links have high rankings. On-page SEO techniques such as page title, header tags, meta tags, target keywords, keywords density, ALT tags, content placement, breadcrumb trail, URL structure and size, internal linking of web pages, site update frequency and site maps play a major role in ranking of a web site. Lalit Kumar et al.[6] lists out the importance of keyword's presence in title tag, meta description about the page's content, meta keyword, heading tab, image alt tag and site map page. Off-page optimization techniques by joining in groups with a link to a site, placing links to sites in social network sites, link building, and blogging. E-Business Toolkit [7] emphasizes keywords of what customers may use, synonyms of key words, keyword selection, research and testing, quality content supporting the overall theme, service location consists of local addresses / city listings, content optimized for social media, blogging , link building, 3-7% of keyword density, short title tag with most important keywords at the beginning, user friendly URL and XML site map. C.Pabitha et al. [8] analyze the importance of keywords in title page, meta description, URL name selection. Reyner D'souza et al. [9] uses document clustering to retrieve relevant pages from a huge collection of search engine result. Hitesh KUMAR SHARMA [10] discusses the role of inbound links ,outbound links and dangling links of a web page. Hema Dubey et al. [11] have tried a normalization technique for improving the page rank of a web site based on mean value of page ranks of all pages of a web site. Ms. Pabitha [12] Discusses eliminating duplicate links to a web site using 301 redirect, rel canonical, no index follow instructions in meta tag, setting preferred domain etc.,[13] statistically says that while ranking web pages google's ranking algorithm gives 23.87% weightage to incoming links of reputed sites to our web site,22.33% weightage is given to link popularity of the content of a web page based on incoming links,20.26% is given to anchor text of links from other sites and social media plays a major role in promoting a site. Usage of keyword research tools, specific keywords help in improving the rank. Relevancy of a keyword, in title tags and headlines, content, content landing page design, bookmarking and sharing also improve the ranking. [14] Speaks about global SEO, does not mean mere translation, need for multi regional, multi lingual web sites,ccTLD, local hosting and local links etc.,[15] says Visibility, openness play a vital role in ranking of a web page [16] Keywords in title tag ,description tag and url, keywords in url must be short and relevant to content,3-7% repetition allowed,breadcrumbs,anchor text for indexing ,images with relevant alt text, heading with proper keywords and freshness of content are the influencing factors in page ranking.[17] lists Lean code, number of links and value of links have a say in ranking.[18] says Rank, authority and relevance, long tail words, good and more content, more number of pages, home page with generic keyword, every product or service with unique page, blogging, every page must be focused, thought provoking headline, bolding keywords, url based on the web page content, picture files names relevancy, headline tags, keyword in anchor text. In off page SEO, reciprocal links, social media and email to spread content and long tail keywords decide on the ranking of a web page. [19] emphases that Only if the content is good ,activities in social media networks about the web site will be more and Google algorithm favors social signals, presence in web 2.0 sites, forum profiles add value, get link from .gov and .edu sites, blog commenting, make videos ,images and ppt and upload to scribd and slideshare, upload content to you tube, quality content may lead to links, on page optimization.[20] reiterates Location of keyword, title tag, keyword density, keyword in url, meta tag, alt txt, anchor text, title length, url length, number of outbound links, link reputation, click popularity.[21] reinforce that domain name and title of web pages should contain the key word for which the web pages are to be optimized, key words within headings, anchor tags, alt-tags and main contents, regular updation of fresh and unique contents, key words more pertinent to the contents, descriptive text for contents ,select keywords judiciously ,keywords in the linking structure , videos,images,audios optimized in similar lines, back linking, social network and groups help in off page optimization. choose reputed web hosting company, web site with static ip address, article distribution to other sites, include forum commenting, submit blog and RSS feed, list your sites in directories, get .gov and .edu links, include social book marking, share web site content on different media regularly and use clear text for links and Usage of tools enhance better ranking of web pages.[22] shows the following methods for increasing the ranking of web sites: links, valuable content, suitable domain, keywords and updation.[23] studies on site optimization with action on code and content of a web site.[24] suggests for web sites of low rank, the following factors: page title, key word optimization, improving link popularity, internal links, join forums and directories and published articles of revered personalities.[25] advocates on-page optimization like page titles, page metadata, headings, body text with relevant content, internal links, breadcrumb trail and image alt tags.[26] indicates that links retrieved using an ecommerce search engine are better than those obtained from most other search engines.[27] results indicate that SEO is an effective method for improving search engine rankings and site traffic.[28] findings indicate that metadata is good mechanism to improve web page visibility. It also states that keywords should come from both title and full-text. Bernard J. Jansen et al.[29] indicates those links retrieved using ecommerce search engines are better than from other search engines. Themistoklis Mavridis et al.[30] finds that the domain of queries affects the web page and the semantic attributes that influence the ranking score in search engines. It also finds that All the SEs still consider link structure as an important characteristic of the web. Gokhan Egria et al. [31] states with competition in online industry Search engine optimization is crucial for keeping the organization in race.[32].applies knowledge based system for ranking web sites.It uses heterogeneous techniques and ontology for knowledge representation for supporting SEO activity. Anindya Ghose et at [33] states that Major Search engine optimization factors influencing web site ranking Hubspot [34] has so many tips with relevant to inbound links generations, Inbound links are called as backlinks that denote how many web sites or web pages link to this particular page. The more the number of backlinks means the page is followed by many web sites. It increases the rank of a web site. Link building can help to increase the inbound link. The more links you have pointing to our site, the more traffic and better rankings our site can secure through search engines. Inbound links bring three major benefits to our business. Social media becomes a key tool in attracting a large number of inbound links at a low cost. Guest blogging can give us access to a new set of viewers and it turn can bring people to our web site. Conducting contests can bring more audience to our web pages. Conducting controversial debate in our web site can attract many users. Publishing in local directories can bring us more people to our web site.

METHODOLOGY

Three search engines were used to find the top 10 web sites. Google being the predominant site was first used. Search through other search engines like bing,yahoo also reveled 90 % similar search results. The following table lists the search results.

Table 1.1

No	Web Site
1	Google.com
2.	Facebook.com
3	YouTube.com
4	Baidu.com
5	Yahoo.com
6	Amazon.com
7	Wikipedia.org
8	qq.com
9	Twitter.com
10	Google.co.in
11	Sina.com.cn

At the second stage keyword “web site analysis” was used to find some of the top Search engine optimization tools.Two tools were identified first one being SEO checkup of site analyzer.com(Tool 1) and the second one being SEO site checkup (Tool 2).

First, Tool1 was applied on all the items of table 1

A data matrix was formed with the following format

$$X_{ij}= \begin{bmatrix} x11, x12....x1j \\ x21, x22...x2j \\ . \\ . \\ xk1, xk2...xkj \end{bmatrix}$$

Where x_{ij} is a value of the j:th variable collected from i:the observation, where i=1,2,….n and j=1,2,….k.

The following table lists the SEO findings for those items

Table 1.2

Web Site	GS	ACCS	DSG	TXT	MM	NW
Facebook.	65	63	77	44	81	56
youtube	61	78	59	58	67	43
baidu	47	57	54	28	88	15
yahoo.com	61	66	76	70	24	52
amazon	51	38	73	36	56	45
wikipedia.org	67	69	75	78	64	44
qq.com	49	54	63	35	57	23
twitter.com	61	57	68	60	67	54
google.co.in	63	60	59	53	74	76
sina.com.cn	43	53	63	35	31	16

The following table shows the correlation between variables chosen

Table 1.3

	GS	ACCS	DSG	TXT	MM	NW
GS	1					
ACCS	0.62	1				
DSG	0.48	-0.04	1			
TXT	0.81	0.64	0.47	1		
MM	0.25	0.11	-0.33	-0.21	1	
NW	0.81	0.22	0.37	0.56	0.16	1

Inference

Table 1.3 reveals the following between different attributes

The following list shows the positive and negative correlation between different attributes
High positive correlation:

Accessibility → Global score
Text → Global score
Network → Global score
Text → Accessibility
Network → Text

Low positive correlation

Multimedia → Global score
Multimedia → Accessibility
Network → Accessibility
Design → Global score
Text → Design
Network → Design
Network → Multimedia

Negative correlation

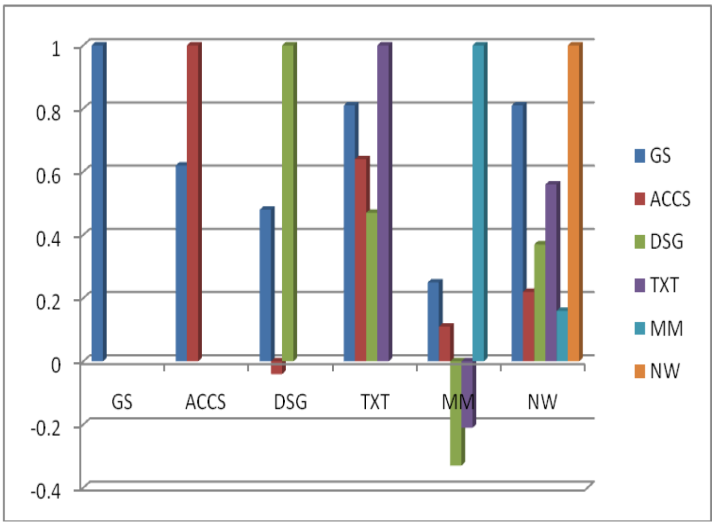
Design → Accessibility
Multimedia → Design
Multimedia → Text

The following are the inferences:

When text content is good global score also increases. And also when network is good global score increases. Good text increases the accessibility of a web site and when accessibility measures are good global score increases. So a transitive relationship is formed between text and global score. It can be inferred that whenever the multimedia content is more in a site less importance is given to text and vice versa. Likewise when design elements are more accessibility becomes more difficult.

RESULTS AND DISCUSSION

Figure 1



The chart 1 shows that a web sites accessibiltiy determines its global score for a score of 0.62. Design determines global score for a score of 0.48.Text determines global score for a score of 0.81.Multimedia has a negative impact of -0.33 on design.Similar scores can be ascertained from the graph.

CONCLUSION AND FUTURE WORK

The above findings reveal that different features of web sites can be analyzed online. But due to time constraints not all online tools were used. An extensive study of various tools may disclose more relevant and authentic information on various web site characteristics

REFERENCES

[1] Elsevier's Getnoticed Promoting your article for maximum impact January 2015;02:4-8

[2] A Review Paper On Web Page Ranking Algorithms Seema Rani , Upasana Garg International Journal Of Engineering And Computer Science ISSN:2319-7242 August, 2014 8:6-11

[3] Ceonex, IncTop Ten Methods to Improve Your Search Engine Rankings 2005:pp 3-12

[4] Google Scholar's Ranking Algorithm: An Introductory Overview, Jöran Beel, Bela Gipp July 2009 : pp 230-241

[5] Exploring SEO techniques for dynamic web sites by wasfa kanwal, *Master Thesis Computer Science,Thesis no: MCS-2011-10 March, 2011 ;6 : 5-12*

[6] SEO Techniques for a Website and its Effectiveness in Context of Google Search Engine, Lalit Kumar and Naresh Kumar International Journal of Computer Sciences and Engineering,vol-2 issue-4 ,2014 ;2-12

[7] Increasing Traffic to Your Website through Search Engine Optimization Techniques, E-Business Toolkit, Ontario.ca/ebusiness july 2011;1: 10-14

[8] Refining Serp – Search Engine Result Page for Enhanced Information Retrieval C.Pabitha , G.Sangeetha, International Journal of Data Mining Techniques and Applications , June 2013 01;3-16

[9] Automatic Link Generation for Search Engine Optimization Reyner D'souza, Apurva Kulkarni, and Imran Ali Mirza, International Journal of Information and Education Technology, Vol. 2, No. 4, August 2012;9:10-16

[10] Enhancing the Ranking of a Web Page in the Ocean of Data Hitesh KUMAR SHARMA, Database Systems Journal vol. IV, no. 3/2013;6:11-16

[11] An Improved Page Rank Algorithm based on Optimized Normalization Technique Hema Dubey Prof. B. N. Roy, International Journal of Computer Science and Information Technologies, , 2011;5:16-21

[12] Search Engine Optimization by Eliminating Duplicate Links Ms. Pabitha. C International Journal of Emerging Research in Management &Technology, January 2015. ;12 :10-16

[13] How to create compelling content that ranks well in search engines by Brian clark founder of copyblogger & scribe pp 104-128

[14] Global Search Engine Optimisation How to speak the language of International SEO, Hallam internet limited;pp-236-256

[15] Ferdinand Budi Kurniawan, Ridwan Sanjaya Search Engine Optimization (SEO) Implementation for Educational Purposes The First International Conference on Interdisciplinary Research and Development, 31 May - 1 June 2011 ;9 :12-17

[16] Fall 2012 Search Engine Optimization: Best Practices for Google Tucker Johnson pp 120-124

[17] Ad Age Insights white paper CJ Newton pp-10-12

[18] Hubspot Introduction to Search Engine Optimization Getting Started With SEO to Achieve Business Goals pp-21-24

[19] A New Methodology for Search Engine Optimization without getting Sandboxed pp-10-14

[20] Search Engine Optimization: A Study Patil Swati P, Pawar B.V and Patil Ajay S Research Journal of Computer and Information Technology Sciences , February (2013);1:10-13

[21] The Foremost Guidelines for Achieving Higher Ranking in Search Results through Search Engine Optimization Khalil ur Rehman and Muhammad Naeem Ahmed Khan, pp-34-26

[22] International Journal of Advanced Science and Technology , March, 2013 ;12 :13-18

[23] Search Engine Optimization based on Effective Factors of Ranking in Web Sites: A Review, Farhad Soleimanian Gharehchopogh, Marjan Mahmoodi Tabrizi, Isa Maleki, pp-120-126

[24] International Journal of Computer & Mathematical Sciences IJCMS ISSN 2347 – 8527 Volume 2, issue 1 January 2014 ;1 :23-28

[25] The Influence Of Google's Ranking Algorithm On Search Engine Optimization (Seo) pp-1-4

[26] Website Design and Development through the Prism of Internet Marketing pp -12-16

[27] Website Usability Evaluation and Search Engine Optimization for Eighty Arab University Websites Ahmad A. Al-Ananbeh*, Belal Abu Ata*, Mohammed Al-Kabi** and Izzat Alsmadi* Aug. 12, 2012;2:10-23

[28] Exploring SEO Techniques for Web 2.0 Websites. Master of Science Thesis in the programme Software Engineering and Technology by Najam Nazar pp 104-109

[29] The effectiveness of Web search engines for retrieving relevant ecommerce links Bernard J. Jansen a,*, Paulo R. Molina b Information Processing and Management 42 (2006) 1075–1098;10:28-23

[30] the Value of search Engine Optimization:An Action research Project at a New E-commerce site pp 1-4

[31] *Ross A. Malaga, Montclair State University, USA,* Journal of Electronic Commerce in Organizations, Volume 5, Issue 3 2007;8;101-106

[32] The impact of metadata implementation on webpage visibility in search engine results (Part II) Jin Zhang, , Alexandra Dimitroff 2005;4:220-225

[33] The effectiveness of Web search engines for retrieving relevant ecommerce links Bernard J. Jansen, Paulo R. Molina 2006;10:20-25

[34] Identifying valid search engine ranking factors in aWeb2.0 and Web3.0 context for building efficient SEO mechanisms Themistoklis Mavridis , AndreasL.Symeonidis , Engineering ApplicationsofArtificial Intelligence 2015;5:13-18

[35] The Role of Search Engine Optimization on Keeping the User onthe Site Gokhan Egria, Coskun Bayrakb, ScienceDirect, Conference Organized by Missouri University of Science and Technology 2014;4:2-13

[36] Knowledge Based System for Intelligent Search Engine Optimization Héctor Oscar Nigro, Leonardo Balduzzi, Ignacio Andrés Cuesta,and Sandra Elizabeth González Císaro Springer-Verlag Berlin Heidelberg 2012 157:4:65-72

[37] Comparing Performance Metrics in Organic Search with Sponsored Search Advertising Anindya Ghose, Sha Yang ADKDD 08Hubspot, social link building August 2008