# Google Page Rank Site Structure Strategies for Marketing Web Pages

**Muneeb Ahmed Farooqi[1], Muhammad Arslan Ashraf[2] and M. Umer Shaukat[1]**

[1]Department of Computer Science, University of Engineering and Technology, Lahore, 54000, Pakistan
[2] Department of Computer Science, National College of Business Administration & Economics Lahore, Pakistan
*Corresponding Author: Muhammad Arslan Ashraf. Email: arslanashraf760@gmail.com

_____

**Abstract:** There are several Search Engines to categorize the web content and show us on base of our search query.These search engines are continuously visiting the pages/sites and gather the information using different techniques called crawling/spidering. On basis of daily content collection all search engines are managing their own indexes for searches.For every business there is a need to make its pages as most top rated/ranked pages by making structurally and content wise batter so that any crawler can easily crawl it and can rank it among the top 10 results.In this thesis only, structural behavior is being discussed in which internal graphical relationships between pages and the loading time of pages in terms of all supportive content .Structural overview includes the HTML tags structure which works as tree flow starting from tag and then moving towards child nodes.Page speed identifies the loading time of a page, it helps search engine to categories pages for mobile devices as well.Sometimes there is a thunder option on mobile search with rank it represents that this page is super-fast in loading. Page loading includes the loading of all content except Ajax base content.According to Google Page, rank is based on page content, page structure and page loading time.As discussion is only Page structure and page loading time so Google already gave some instructions related to Page structure and page loading speed but those instructions are not enough also it need to discover the new dimensions to make business pages among the top-rated results.

**Keywords:** Crawling/spidering; Ranked pages; HTML tags; Child nodes; Google Page;

## 1. Introduction

There is another perspective when search engines showing our pages and content in form of search results. In this case what search engines are doing, they simply crawl our site and get the information they needed and then compare this information with others site's relative data to validate quality and plagiarism of content and then show it in searchers. The algorithm Google is using to rank the data is called Page Rank. Every search engine is using its own algorithm with different name. That is a very fruitful case where our page comes on top easily just because Google get information easily but if that information is not visible to crawler which is coming on our site to scrap the data them how is it possible that it can compare our content with others without having it. What it needs to do is think like search engines, how they are coming on sites, how they are getting the information, how they are categorizing the content, how they are ranking our material. Currently there are many SEO techniques to boost our pages also having many search engines guides but these are not enough because search engines are continuously upgrading their rankings algorithms and SEO tools are not upgrading in same manners. According to GOOGLE Page Rank is using web content and web structure to rank search results, here Page structure is important.

- Page Header Layout
- Tags information
- Page inner graph relationship
- Page bidirectional root
- Page size in form of content
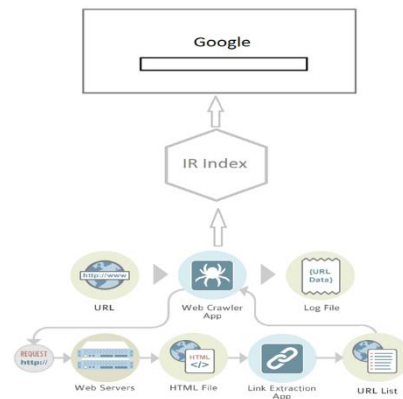- Page orientation of defined tags
- Page loading time



**Figure 1: Crawling Struture**

In Figure 1 Web crawler is fetching the html then passing it to web scrapper, scrapper is removing the extra text from html file, and then getting required results (links) these are the links system is using to navigate the next results. This figure is explaining the parsing of Google search results. After getting the required URLS crawler pass it to some file or storage machine. After all this lengthy process Page Rank works and performs some calculation to validate the content if content is valid then it Perform some Information Retrieval techniques to index the data for fast retrieval

Google can show 100 results against each term and each result contain a website so at end perform 70000 calculations on web pages. Sample of some keywords are given.

**Table 1: Marketing Terms**

| Category | Terms |
|----------|-------|
| Marketing | Online Sale |
| Marketing | Ecommerce |
| Marketing | Buy online |
| Marketing | Online Cloths |
| Marketing | Online Shopping |

Containing of 70000 web pages with 70000Jsons records, also 700 Google Search Results HTMLS and JSONS. These HTMLS and JSONS will be uploaded if needed. To parse the documents system used **NOKOGIRI** Technique. Then find the differences of top rank sites and least rank sites. Some of the data is recently fetched but most of the old data is gathered from private SEO organizations. Then figure out some points on which Google is ranking.

**Table 2: Crawling Counts**

| ITEM | Count |
| --- | --- |
| Total Google Search Pages | 700 |
| Results in one page | 100 |
| Total Results | 70000 |
| Total Web pages | 70000 |
| Total interlinked Web pages | 11200 |
| Top Ranked pages | 7000 |
| Least Ranked Pages | 5691 |

Each keyword was passed to Google and google returned SERP. SERP is the results page that any search engine shows against any regular text search which includes ranks and SFTS. According to user's perspective Top ranked results are google top 10 results got 7000 top tanked results and 5691 least ranked pages. Least ranked pages are less because some of the keywords were not having 100 results. All Structural and content loading techniques were applied on this site and observer 15 days. Google took 15 days to make it up rank. After 15 days this site was in top 5 results in global organic results and in top 3 sites in US. Paper used some Japanese keywords to find out the linguistic difference effect on Page Rank but ranks remains same.

## 1.1. Problem Statement

Search engines are continuously visiting the pages/sites and gathering the information called crawling/spidering. Also make them simple to compare with other pages which will help the search engines to index it properly. Although search engines are having instructions to make pages better in means of page structure and page loading speed but these instructions are not enough to create a generic top-rated page because Google Page rank is changing its algorithms regularly. It needs to get common points that Google Page rank always considered. Our main problem is to find out the Layout Google is using internally to evaluate the sites and speed loading mechanism because given structure in not enough.

## 1.2. Research Objectives

We need to think like search engines, how they are coming on sites, how they are getting the information, how they are categorizing the things, how they are ranking our material. Currently have many SEO techniques to boost our pages also have many search engines guides. The main objective of this thesis that need to make these pages So here are looking only structural perspective of a page and internal graphical relationships between pages. Using these relationships, can define the hierarchy of a site same as search engines Page Rank is an algorithm used by Google Search to rank web pages.

- To investigate the existing SEO tools for structural behavior of a webpage in means of marketing Web pages.
- To arrange a list of behavioral patterns for Google Page Rank to predict the bottom-line ranking strategies.
- To make web pages as most top rated/ranked pages by making structurally batter so that any crawler can easily crawl it and can rank it among the top 10 pages. Also make them simple to compare with other pages which will help the search engines to index it properly.
- To study and analyze that which changes are required in existing structure of Google web page ranking.

## 1.3. Research Questions

We will try to answer the following questions:

- What is the Layout Google is using internally to evaluate the sites?
- What are the required changes in existing structure of Google web page ranking?
- What is the structural behavior of top marketing Web pages?
- What are the parameters to ensure the accuracy of SEO masters?

## 2. Literature Review

All search engines are crawling billions of pages on daily basis, there is already a huge work on all aspects including how to request the web page, how to open the page, how much delay needed, how much content need to filter, how frequent can call a page, how frequent data is changing on page, what are the rules to fetch the data, which data is useful, what technique can use to get accurate results, how to compare these results with others, how to reduce plagiarism. How to identify which data is authentic and belongs to author after all of that how to rank the data and what user is expecting against the search query .Some pages are showing static data, just hit the URL and data will be displayed, some are having security measurements in case they have IP restriction which validate if requested IP is dynamic or not.

In all this validation pages takes time and the simple get request won't work here to add delays, use browser strategies or selenium which manipulate DOM and made changes that server things requested host is an human and show the data or add delay after requesting the page and after a random seconds of time request to verification of data. When each request will go with a different IP address and server won't able to identify the request source frequency. But then another term will come where server will able to check the content type of each request if all content requests belong to same domain the most of the chances are user is using proxies. Google is using Page Rank for results ranking mechanism, according to Google "Page Rank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. On user search query request came to a centralized system which is also using a small part of Page Rank and render a SERP page. Search Engine Ranking Report is the page that any search engine shows against a search.
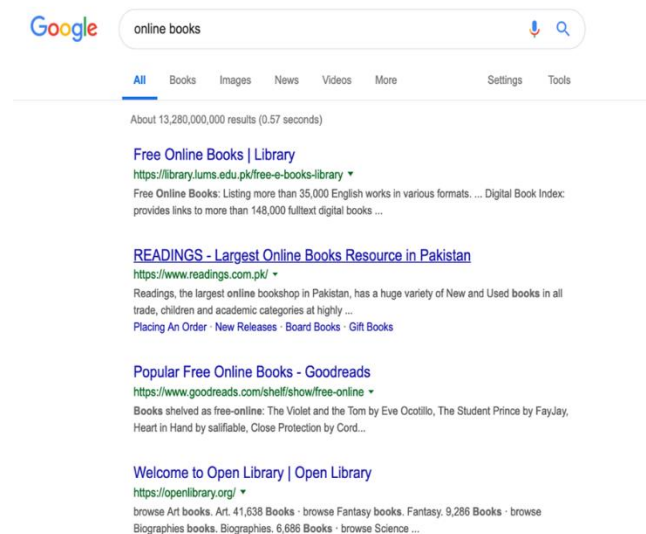


**Figure 2: SERP**

A simple example of SERP is given where "Online Books" is the user query, under search bar google shows results count stats "13,280,000,000" (Kelly & Nixon, 2013). It means google is having that number of ranked results against our query. Under this blue color shoes, the title of ranked results. Sometimes SERP may contain properties, it's the peak point of SEO if any search engine is showing you in properties. Google is using user's data / expected reviews and comments to show data into properties. SERP feature Trend is a Result on Search Engine Results Page (SERP) that is not an organic (Haddow & Genoni, 2010) result. Google is having 4 types of SERP 341 feature trends.

   **1) Rich Snippets**

      User expected behavior, mostly works for online shopping and hotel reviews result (e.g., review stars for product ratings)(Andrikopoulos, Sun, & Guo, 2017)

   **2) Paid Results**

That is the part of Google Webmaster tool where google provides facility to purchase keywords are use them on your webpage against a paid rank. (e.g., AdWords or Google Shopping)(Andrikopoulos et al., 2017)

**3) Universal Results**

These are additional organic results. (e.g., image results, new results, featured snippets)(Andrikopoulos et al., 2017)

**4) Knowledge Graph data**

All boxes except ranks are coming into this section (e.g., weather, Celebrity Knowledge Panel)(Andrikopoulos et al., 2017)

In the early days of Google, every result on the SERP looked the same, and these traditional organic results have not changed much in appearance over the years.
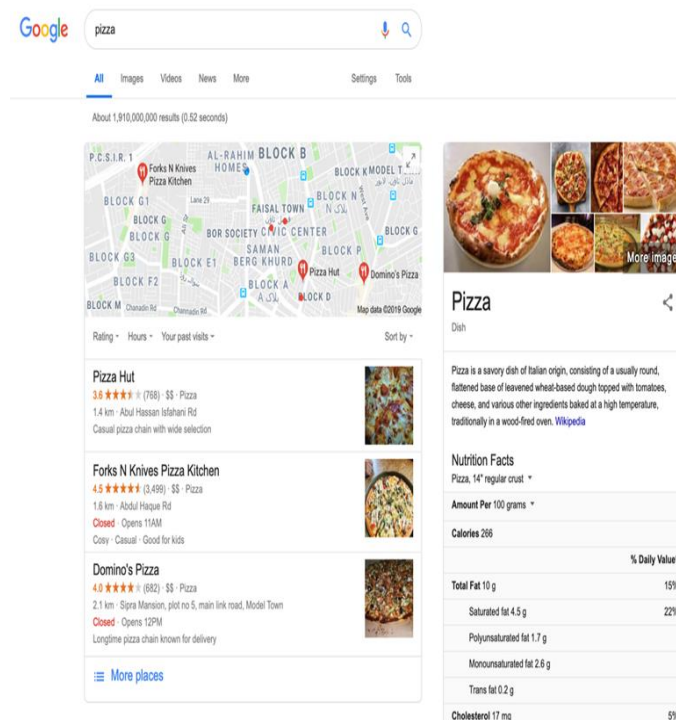


**Figure 3: Location base search**

Here in image SERP is including different types of feature elements. Given SERP includes map location, knowledge panel and rating panel. However, beginning with AdWords (Andrikopoulos et al., 2017) (Haddow & Genoni, 2010) in 2001, Google has been very busy adding other non-organic results into the SERP(Vargiu & Urru, 2013). Organic Results are the main section of search page. The part which shows ranked pages is called organic results. Page Rank first focus on organic ranks because these are the expected results for a user. Till now google introduced 26 types of SERP properties.

**Ads on Top**

Ads are the part from where search engines got most of their income. Whenever do some search, Google shows some ads on top of the results as rank result. Google differentiate it with an Ad check in green color (Andrikopoulos et al., 2017) Here in this search query first 3 ranks are shown as "ads on top" after this Google shows the original organic result.
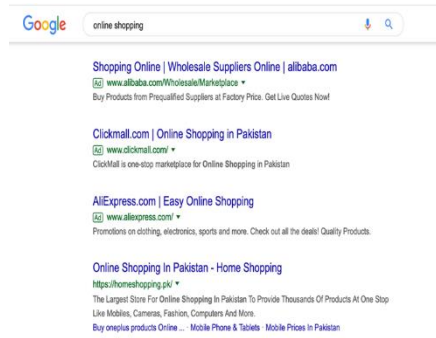
**Figure 4: Ads on Top**

**Ads on Bottom**

Google shows ads in bottom with least matching but paid content. Why it shows ads on bottom, because these are the part of their agreement that if someone is choosing the monthly strategy from Google then this company will give some extra credits to that person and after one-month ads will go into bottom as bonus.(Andrikopoulos et al., 2017)
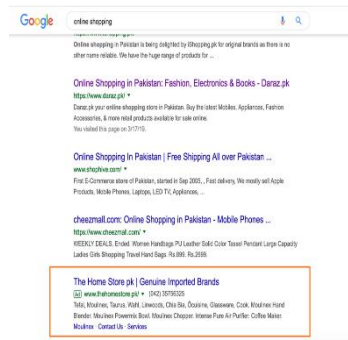


**Figure 5: Ads on Botom**

- **Sports**
- **Info**
- **Disambiguation**
- **General converter**
- **Related questions**
- **Currency Conversion**
- **Translation**
- **Dictionary**
- **Stock**
- **Nutrition**
- **Weather**
- **Ads on sidebar**
- **Local pack**
- **Knowledge panel**
- **Image links**
- **Video links**
- **Shopping**
- **News**
- **Answer box**
- **Google flights**
- **Home service ads**
- **Google Jobs**

These are all properties that Google introduced(Andrikopoulos et al., 2017) till now. Some of them are regional and some on them are universal. Like home service adds and shopping is only active in America, Canada, Australia and European countries. Other hand Google Jobs is universally active. Google is also using each property as a separate tool (Deng & Ong, 2014). Starting from 1st property. Google is using Ads as a paid service in Google webmaster (Deng & Ong, 2014)tool and also including it into SERP. Local Pack and knowledge panels are the part of review ever given on any site. Google gather that reviews data from those sites and shows in SERP that for a specific webpage what other people thinks. Image links and video links are coming from social sites and YouTube. User upload their images there and Google crawl them and show them as image and video tabs.

Shopping(Haugh, 2007) the new feature that Google introduced recently, it's a regional base feature which is only active in few countries (Jiang, Hu, & Li, 2009). Google is having an entirely separate tab as well as part of SERP results.



**Figure 6: Shoping screen**

According to dataset are using for Google SERPS have these stats of properties occurrence. (Zhao, Lu, & Duan, 2009)



**Figure 7: Statistics**

To scrap the data most important factor is the technique, which technique are using to fetch the data. Sometimes our ultimate goal is to analyses the data or perform some machine learning to observe the meaning of unstructured data .So, in these scenarios will first fetch the raw data, perform some categorization, make them in structures way and then apply our techniques But sometime Appling techniques process comes before structuring the data because most of the time are using techniques to convert data from unstructured to structured form .

After fetching and converting the data the most important step is to analyzing the data, to analyses the data Google is using different approach the major objectives are to identify data is plagiarized or not, information is valid, related information, related pages, relations between the similar data pages and many more Google is using to build Artificial Intelligence related data heuristics. According to Google Lighthouse have several properties which takes an important role to boost a page performance. These are the core part of any search engine.

- Performance
- Progressive web application
- Accessibility
- Best Practices
- SEO

### 2.1. Performance

It represents the performance of a webpage. To measure the performance, have different techniques and tools. According to Google research association performance takes 20% in ranks of an organic result builder. It contains 5 sub properties.

- First Paint Concertful
- First Meaningful Paint
- Speed Index
- Interactive
- First CPU Idle

| # | Property | Weight | Group |
|---|----------|--------|-------|
| 1 | First Contentful Paint | 3 | metrics |
| 2 | First Meaningful Paint | 1 | metrics |
| 3 | Speed Index | 4 | metrics |
| 4 | Interactive | 5 | metrics |
| 5 | First Cpu Idle | 2 | metrics |

**Figure 8: Performance Matrix**

This table explains the weight of each property with group association(Zhao et al., 2009). Starting from first one. First Content Paint is a term to determine what is the very first user interface frame a web page show. To check its appearance, can use TESTALIO. First system refreshes the page and pause it is refreshing states(Kelly & Nixon, 2013)to know how first frame looks like. Let's get an example of DEMANDSPHERE official site.



**Figure 9: Loading Frames**

### 2.2. **Progressive Web Application**
It's mainly related to network calls. Which determines 12 different statistics. Each statistic is having its own weight in Page Rank calculation.

| # | Property | Weight | Group |
|---|----------|--------|-------|
| 1 | Load Fast Enough For Pwa | 7 | pwa-fast-reliable |
| 2 | Works Offline | 5 | pwa-fast-reliable |
| 3 | Offline Start Url | 1 | pwa-fast-reliable |
| 4 | Is On Https | 2 | pwa-installable |
| 5 | Service Worker | 1 | pwa-installable |
| 6 | Installable Manifest | 2 | pwa-installable |
| 7 | Redirects Http | 2 | pwa-optimized |
| 8 | Splash Screen | 1 | pwa-optimized |
| 9 | Themed Omnibox | 1 | pwa-optimized |
| 10 | Content Width | 1 | pwa-optimized |
| 11 | Viewport | 2 | pwa-optimized |

*Figure 10: PWA*

In 2nd property "Works Offline" means incase network got stuck or not having any connection is still page loading, yes, it's confusing term but if webpage is having offline extension added, one a page is opened in user browser. Browser will save it permanently and incase next time don't have any connection browser will check the history and stored pages and will show it to user but it has a negative effect as well that all pages will store on use rend so after a long searching user will have a very high data on its system disk Next point is "Service Worker". It defines if a server is having spare workers to run its services. Sometimes have notifications to enable, these are coming from these service workers.

- redirect from http to https
- Redirect from https to http

Like if use NGINX then can write it in NGINX configurations that incase a user request with http header, then simply redirect this user to https.2nd aspect in which a webpage which hosts on a https network is having links of a http page. As http will have lower rank in page rank so if a higher rank will refer a lower rank, it'll cause rank issue for higher rank bit for lower rank it'll be a plus point because higher rank is giving points to lower rank. JAVASCRIPT is widely used as use rend framework which means it allows systems to use user's system rather than putting load on server simply use end user's system to render / populate some data. It's having a very low weight in Page Rank but still google check either a server is completely sending load to user system or balancing some of oy on its end or not.

## 2.3. SEO

SEO is the most important term to rank a webpage. According to Google it contains 12 different properties. Each property owns its own group with a unique weight. A full detailed table is given;

| # | Property | Weight | Group |
|---|----------|--------|-------|
| 1 | Viewport | 1 | seo-mobile |
| 2 | Document Title | 1 | seo-content |
| 3 | Meta Description | 1 | seo-content |
| 4 | Http Status Code | 1 | seo-crawl |
| 5 | Link Text | 1 | seo-content |
| 6 | Is Crawlable | 1 | seo-crawl |
| 7 | Robots Txt | 1 | seo-crawl |
| 8 | Hreflang | 1 | seo-content |
| 9 | Canonical | 1 | seo-content |
| 10 | Font Size | 1 | seo-mobile |
| 11 | Plugins | 1 | seo-content |
| 12 | Tap Targets | 1 | seo-mobile |

**Figure 11: SEO**

Mostly SEO lies in 3 different groups(Arlitsch & Obrien, 2013)
- SEO MOBILE
- SEO CONTENT
- SEO CRAWL

According to table starting from first property. "VIEWREPORT" is a high-level site map which a system put with extra headers so that during mobile size conversion system can understand the pages hierarchy and resolution sizes. 2nd part is "DOCUMENT TITLE", it's the text always see in first line of any result in blue color. So, whenever do some search our search engine first of all check all the documents titles if it exists in it. It's for search engines and search engines use term technique so better is to use keywords rather than a plain English paragraph it gives us many benefits like in a very short terms or lines can engage many categories just writing their names. It's very important to choose links text very carefully rather than to write simple keywords like "Click here". Search engines. mostly compares links text with URL description or title so whatever text you choose it must be according to that URL description on title. Next point is "IS CRAWLABLE", it means to check if a webpage is crawl able is not it's very important in case a webpage is not crawl able then no search engine will crawl it and in result that page will never be the part of organic results.

**3. Research&Methodology**

In this section discuss the methodology used to perform our research on Page structure strategies. It'll include all the process from Data set collection to analyzation so can categorize in 3 parts and then sub parts. Starting from dataset preparation where gather marketing keywords using Google console, DE-MANDSPHERE, HREF and TESTALIO APIS. These are all the tools which provides search engine optimization-based keywords and competitors scores. After gathering the keywords next step is to collect organic results from Google for given keywords set but that is not an easy job because by default Google provides 10 results per page and if do it programmatically Google block us just after few requests so better option is to use selenium base systems to show google that a human is operating the system, it's a length but useful process by doing this write a fetcher.

**3.1. Dataset Preparation**

Marketing Keywords collection
Google Organic Ranks collection for marketing keywords
Crawl Google Pages and get ranked pages URLS
Crawl given URLS and get site map
Gather data from site map and transform in JSON format
Apply both Google Lighthouse and Custom algorithms

**3.2. Dataset Authenticity and Verification**

Pass the data to custom algorithm
Pass the data to Google Lighthouse
Pass the data to DEMANDSPHERE AND HREF
Compare results

**3.3. Evolution Phase**

Add custom attributes to webpage
Add Google attributes to webpage
Search Marketing Keywords on Google for organic results after 15 days
Parse the html and gather ranking data
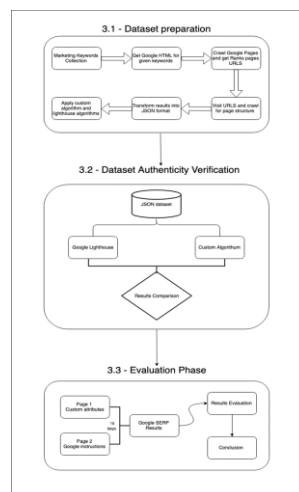Analyze results by comparing with initial results.



**Figure 12: Methodology**

**3.4. Dataset Preparation**

Google Search Console (Arlitsch & Obrien, 2013) (Chen, Chiang, & Storey, 2012)
DEMANDSPHERE
MOZ SEO Tool

After gathering 700 marketing keywords, Google queries for all these keywords were prepared so that can search them on google. But issue was on Google search results. If call google search query pro-grammatically then Google treats it as thread and always show captcha. But issue still remains, after 60-70 request google shows captcha again So now I used proxies which make a perfect solution to gather all SERP one by one. Usually, Google shows 10 results per page but as customized queries were used by passing number=100 PARAM so Google starts showing 100 results on per page. Here is a thing to note that these HTMLS are only for Google Desktop. But first need to validate data efficiency So used Lighthouse and Google Console for validating data efficiency. Data efficiency report is given in screenshot:
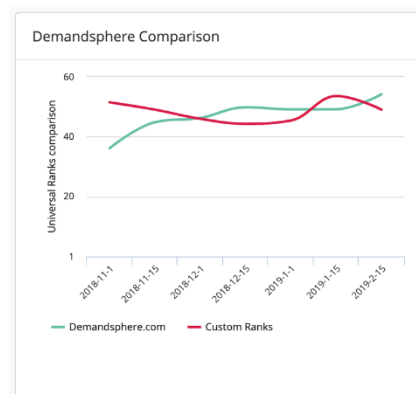


**Figure 13: Ranks Comparision**

After data validation have added different attributes into the Web pages so that can test them. Attrib-ute details are given.

- Page title length(Vaughan, Shaw, & Technology, 2005)
- Eliminate render/blocking resources from webpage(Vaughan et al., 2005)
- Minify CSS and compress in SHA formats(Vaughan et al., 2005)
- Defer offscreen / SEO SERP images with meta tags(Vaughan et al., 2005)
- Enable text / assets compression(Vaughan et al., 2005)
- Avoid multilevel webpage redirects / renders(Vaughan et al., 2005)
- Avoids an excessive / larger DOM size (page size)(Vaughan et al., 2005)
- Minimize Critical Requests Depth and stack level should be less than 11 stokes(Vaughan et al., 2005)
- JS execution time once request rendered(Jiang et al., 2009)
- Never respond with a http 200 status when offline(Vaughan et al., 2005)
- Page transitions / firewall don't feel like they block on the network(Vaughan et al., 2005)
- The page contains a heading / title, skip link, or landmark region / location(Vaughan et al., 2005)
- Avoids `document. write () `, it never be in DOM(Vaughan et al., 2005)
- Meta SEO tags with proper baseline(Vaughan et al., 2005)

As added Google Analytics in it so that can have all stats when and with what state Google crawled it. Go according to gathered information Google crawled it after 2 days and after first crawl updated the content not structure so that in next attempt Google can schedule an early crawl. So, in 2nd term Google crawled the page after 4 days and set time of 12 hours. And after that Google starts crawling the page after 12 hours. After comparing with first 2 web pages, it was analyses that their content was much better as content is not part of this research so can negotiate that factor.

## 4. Proposed model, Framework and Solution

Global web means the area were most of the user's search relies and trust for the most rated results organized by these search companies, so these companies always focus to improve quality of their content

and use better SEO related techniques so that they can come in first 10 results of any search engine. In the global list of search providers Google is on top. According to 2018 stats 93% web search was performed by Google search Engine in last 5 years, that makes Google the best choice for people. That is not the only thing, Google also have a very responsive search mechanism which response with affective ranked results. There are other search ranks provider as well and they have a very targeted area like;

### 4.1. Yandex

Russian base search engine which provides only Russian localization also, it's the most common linguistic Application programming interface provider which owns even more than Google market for its language translation techniques just because of its accuracy. Other than this it provides free business email service but still Google is on top in this service because Google provides scam detection interface for unregistered email or unauthorized IPS coming emails. Yandex only shows the organic results and not having any feature properties that's also a reason it's not being so popular globally.
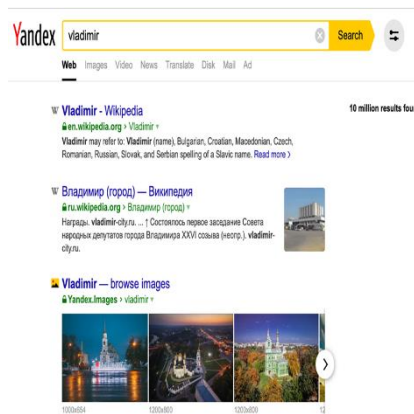


**Figure 14: Yandex**

### 4.2. Baidu

Baidu is a Japanese base search engine which provides mostly Japanese results. Most of the Japanese people relies on this search engine just to maintain their culture.



**Figure 15: Baidu**

### 4.3. Yahoo

Yahoo is a globally famed search engine but its fame is not because of its searching capabilities. Meanwhile Google was so strong that people were not using yahoo for shopping on search mechanism. Google took advantage of yahoo slowness and take the maximum market. Not only this Google introduced new feature in SERP. Other than this Google starts its micro services and add them into search results and today more than 22 micro services are running in parallel with organic results. (Organic results are the results those come by in result of a normal user search on a search engine
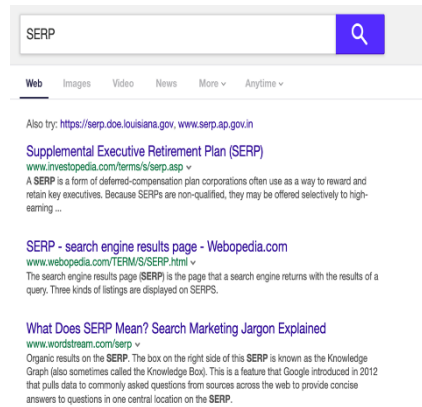
**Figure 16: Yahoo**

**4.4 Google**

Google is the most famous and top-rated search engine in market. According to global stats in 2018 93% global traffic was redirected from Google so can say Google knows about every brand, users need and all the stuff that can be gathered by user's data. So, Google earns the profit from these user's data. Google provide this data on public terms and sale it to marketers and big companies so that they can expand their businesses. To do this BUSINESS / MARKETING persons pay some amount to Google. Like Amazon is having different stores so in case someone search AMAZON on Google first I'll get the 2 paid results if they still are paying Google to make their site on top and how Google works is, Google make the results on top and add a add sign that this content is paid, but most of the users are not aware of this thing and in result of their query they always get that paid content and once they clicked on that content Google will cause a simplest amount to its owner, it means Google redirects the traffic to that particular webpage, recently Google announced that Google is starting Google Shopping feature which later it neglects after dictation with Amazon that Google will not launch stores like Amazon.
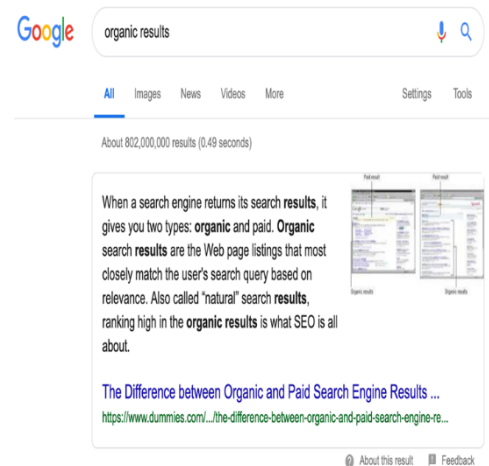


**Figure 18: Google Ads**



**Figure 17: Organic Results**

But google can show the users to search results so on initial level Google started this service on a very small scale starting with US and after getting better response move it to central Europe. Still Asia is not having this service but most of the European countries, Australia, Canada and USA is using Google Shopping for shopping related searches. So, how google actually works for Google Shopping is using affiliate marketing, Google is not crawling like other universal search results. Google shows Google shopping results in Organic results as well as in Google Shopping tab. So, Google divide these results in 2 types.

- Organic search/type 1
- Google search/type 2

### 4.5. Ruby and Python

Ruby on Python both are backend languages to operate datasets. Ruby and Python both are widely used to manage datasets, both have almost the same syntax but Python is having advantage in few fields over Ruby. Ruby is having an edge over ruby is that it can map the results. Pure Python is not having that functionalities, yes it has few libraries but all of them are using GO to map datasets. Like if want to draw a R graph, Python won't be able to help us directly, it'll refer libraries and all of its graphical libraries are using GO underline so whenever use some third-party tool for measurements it's possible it shows some delays, might have some dependency issues and also data integrity can be compromised as well.

| RUBY | PYTHON |
|---|---|
| • More Magical<br>• Created in 1995 by Yukihiro Matsumoto | • More Direct<br>• Created in 1991 by Guido Van Rossum |
| • Tons of features out of the box for web development<br>• Quick to embrace new things | • Very easy to learn<br>• Diverse community with big ties to Linux and academia |
| • Can be very hard to debug at times | • Often very explicit and inelegant to read |
| • Ruby on Rails - Started in 2005 by David Heinemeier Hansson | • Django - Started in 2003 by Adrian Holovaty and Simon Willison |

**Figure 19: Ruby vs Python**

### 4.6. Ruby on Rails

Ruby on Rails is a ruby framework was designed back in 2004. I choose this framework because this is one of the best frameworks to operate data-oriented research projects and I was also having core knowledge of this frameworks. It provides many built-in functionalities and underline gems which helps to configure all the system just in few minutes.



**Figure 20: Rails**

### 4.7. NOKOGIRI

NOKOGIRI is one of the widely known crawling and parsing framework, its name was chosen by name of its founder NOKOGIRI OHARA which was originally from Japan and also in one of the founders of ROR. So how NOKOGIRI works is, it provides its own language of working. First it converts all the content to its designed structure and then anyone can transform it into xml or another basic stream line. These are the basic parts to work in it

- Crawl webpage
- Convert webpage content into NOKOGIRI format
- Nodes and Method
- Selenium with Capybara
- MECHANIZER
- Proxy crawl
- Google lighthouse

| # | Property | Weight | Group |
|---|----------|--------|-------|
| 1 | First Contentful Paint | 3 | metrics |
| 2 | First Meaningful Paint | 1 | metrics |
| 3 | Speed Index | 4 | metrics |
| 4 | Interactive | 5 | metrics |
| 5 | First Cpu Idle | 2 | metrics |

**Figure 211: Performance Matrix**

| # | Property | Weight | Group |
|---|----------|--------|-------|
| 1 | Load Fast Enough For Pwa | 7 | pwa-fast-reliable |
| 2 | Works Offline | 5 | pwa-fast-reliable |
| 3 | Offline Start Url | 1 | pwa-fast-reliable |
| 4 | Is On Https | 2 | pwa-installable |
| 5 | Service Worker | 1 | pwa-installable |
| 6 | Installable Manifest | 2 | pwa-installable |
| 7 | Redirects Http | 2 | pwa-optimized |
| 8 | Splash Screen | 1 | pwa-optimized |
| 9 | Themed Omnibox | 1 | pwa-optimized |
| 10 | Content Width | 1 | pwa-optimized |
| 11 | Viewport | 2 | pwa-optimized |
| 12 | Without Javascript | 1 | pwa-optimized |

**Figure 22: Progressve web aplication**

| # | Property | Weight | Group |
|---|----------|--------|-------|
| 1 | Aria Allowed Attr | 3 | a11y-aria |
| 2 | Aria Required Attr | 2 | a11y-aria |
| 3 | Aria Required Children | 5 | a11y-aria |
| 4 | Aria Required Parent | 2 | a11y-aria |
| 5 | Aria Roles | 3 | a11y-aria |
| 6 | Aria Valid Attr Value | 2 | a11y-aria |
| 7 | Aria Valid Attr | 5 | a11y-aria |
| 8 | Button Name | 10 | a11y-names-labels |
| 9 | Bypass | 10 | a11y-navigation |
| 10 | Color Contrast | 6 | a11y-color-contrast |
| 11 | Document Title | 2 | a11y-names-labels |
| 12 | Duplicate Id | 5 | a11y-best-practices |
| 13 | Html Has Lang | 4 | a11y-language |
| 14 | Html Lang Valid | 1 | a11y-language |
| 15 | Image Alt | 8 | a11y-names-labels |
| 16 | Label | 10 | a11y-names-labels |

**Figure23: Accesibility**

| # | Property | Weight | Group |
|---|----------|--------|-------|
| 1 | Appcache Manifest | 1 | |
| 2 | Is On Https | 1 | |
| 3 | Uses Http2 | 1 | |
| 4 | Uses Passive Event Listeners | 1 | |
| 5 | No Document Write | 1 | |
| 6 | External Anchors Use Rel Noopener | 1 | |
| 7 | Geolocation On Start | 1 | |
| 8 | Doctype | 1 | |
| 9 | No Vulnerable Libraries | 1 | |
| 10 | Notification On Start | 1 | |
| 11 | Deprecations | 1 | |
| 12 | Password Inputs Can Be Pasted Into | 1 | |
| 13 | Errors In Console | 1 | |
| 14 | Image Aspect Ratio | 1 | |

**Figure24: Best Practices**

| # | Property | Weight | Group |
|---|----------|--------|-------|
| 1 | Viewport | 1 | seo-mobile |
| 2 | Document Title | 1 | seo-content |
| 3 | Meta Description | 1 | seo-content |
| 4 | Http Status Code | 1 | seo-crawl |
| 5 | Link Text | 1 | seo-content |
| 6 | Is Crawlable | 1 | seo-crawl |
| 7 | Robots Txt | 1 | seo-crawl |
| 8 | Hreflang | 1 | seo-content |
| 9 | Font Size | 1 | seo-mobile |
| 10 | Plugins | 1 | seo-content |
| 11 | Tap Targets | 1 | seo-mobile |

**Figure25: SEO**

## 5.Results and Evaluation

In this section authors discuss all the discovered results and their evaluations using 3 different parties. As our all research is evolving around Google so first of all will use Google Lighthouse to evaluate our results. Then we'll compare our results with DEMANDSPHERE and TESTALIO by customizing their search properties.

### 5.1. Findings

The major findings are given below;

Use LSI (Long Semantic Indexing) Keywords (limit < 30)

Header sections must be unique and secure using CSRF token or sha256 hash

Page description should not exceed from 300 characters

Page must have a customize meta tag

http status 200 is the best for each increment from 200 will decrease it in points in Page Rank

Optimize the webpage for voice search.

Use Bullets rather than to create 2-3 lines of paragraph.

Never load content with page if webpage is dynamic.

Network load

Stop loading extra content

Save user time

Boost webpage CTR (Click Through Rate).
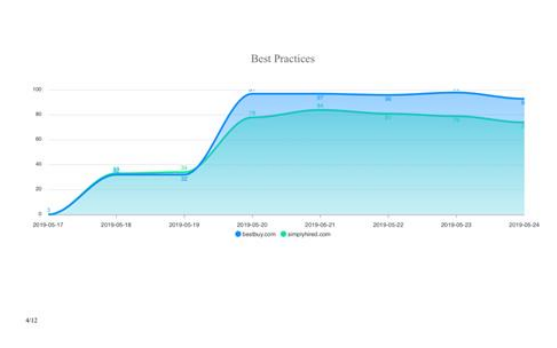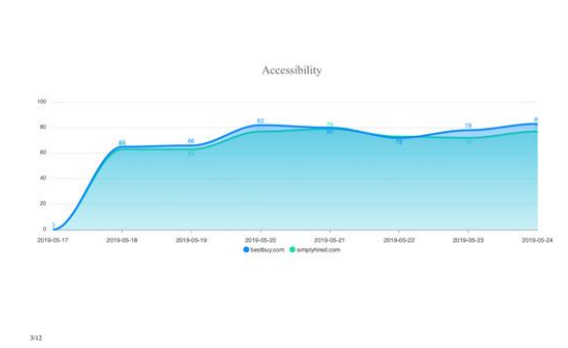
Include Google Properties in webpage structure.

### 5.2. Results

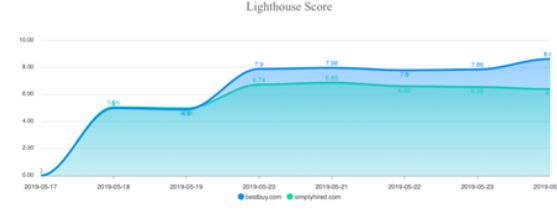The computed results are shown in below figure that are used in the google page rank site structure for web pages.
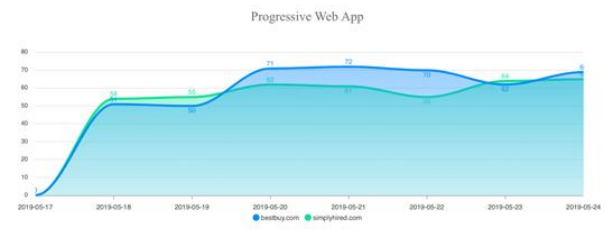
### 6. Conclusion and Future work

In this section we'll discuss what have achieved in this experimental research and what can be achieve in future work also where can use this study. This research is completely for marketing Web pages. Yes, some of its factors can be used for related domains but still it's very targeted. can enhance it and create a globalize system where by putting terms of any field can generate the useful results. As Google is using targeted techniques so need to make sure to use GEO location as a compulsory parameter in search query. About the market, can use it in marketing field. Today internet is the global market of each business. don't know the actual size of Amazon store but have a really big image about it just because whenever do search something on Google Amazon or Alibaba came to the top also have heard news about these big names so that makes our mind about their hugeness. So now in coming future if any business needs to expand it globally and for this every business need search engine optimization.

### References

1. Andrikopoulos, P., Sun, J., & Guo, J. J. I. J. o. M. F. (2017). Ownership structure and the choice of SEO issue method in the UK.

2. Arlitsch, K., & Obrien, P. (2013). *Improving the visibility and use of digital repositories through SEO: A LITA guide*: American Library Association.

3. Barber, B. M., Lehavy, R., & Trueman, B. J. J. o. f. e. (2007). Comparing the stock recommendation performance of investment banks and independent research firms. *85*(2), 490-517.

4. Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. J. B. (2009). Citizen science: a developing tool for expanding science knowledge and scientific literacy. *59*(11), 977-984.

5. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., . . . Wiener, J. J. C. n. (2000). Graph structure in the web. *33*(1-6), 309-320.

6. Byun, J. H., Ahn, J. B., Kim, S. Y., Kang, J. H., Zang, D. Y., Kang, S. Y., . . . Kim, B.-S. J. T. K. j. o. i. m. (2019). The impact of primary tumor location in patients with metastatic colorectal cancer: a Korean Cancer Study Group CO12-04 study. *34*(1), 165.

7. Cai, D., Yu, S., Wen, J.-R., & Ma, W.-Y. (2003). *Extracting content structure for web pages based on visual representation.* Paper presented at the Asia-Pacific Web Conference.

8. Capstaff, J., Fletcher, J. J. J. o. B. F., & Accounting. (2011). Long term performance and choice of SEO method by UK firms. *38*(9-10), 1262-1289.

9. Chen, H., Chiang, R. H., & Storey, V. C. J. M. q. (2012). Business intelligence and analytics: From big data to big impact. 1165-1188.

10. Deng, X., & Ong, S. E. (2014). *Real earnings management, liquidity and REITs SEO Dynamics.* Paper presented at the AREUEA-ASSA Annual Meeting.

11. Ding, Y., Yan, E., Frazho, A., Caverlee, J. J. J. o. t. A. S. f. I. S., & Technology. (2009). PageRank for ranking authors in co-citation networks. *60*(11), 2229-2243.

12. Dos Santos, M. B. J. T., & Talk. (1996). The textual organization of research paper abstracts in applied linguistics. *16*(4), 481-500.

13. Du, W., Zhong, W., Tang, Y., Du, W., & Jin, Y. J. I. t. o. i. i. (2018). High-dimensional robust multi-objective optimization for order scheduling: A decision variable classification approach. *15*(1), 293-304.

14. Fanelli, D. J. P. o. (2010). Do pressures to publish increase scientists' bias? An empirical support from US States Data. *5*(4), e10271.

15. Franceschet, M. J. I. P., & Management. (2010). Ten good reasons to use the Eigenfactor™ metrics. *46*(5), 555-558.

16. Ginde, G., Saha, S., Mathur, A., Vamsi, H., Dey, S. R., & Gambhire, S. S. J. a. p. a. (2018). Use of NoSQL database and visualization techniques to analyze massive scholarly article data from journals.

17.   Ginglinger, E., Matsoukis, L., Riva, F. J. J. o. B. F., & Accounting. (2013). Seasoned equity offerings: Stock market liquidity and the rights offer paradox. *40*(1-2), 215-238.

18.   Haddow, G., & Genoni, P. J. S. (2010). Citation analysis and peer ranking of Australian social science journals. *85*(2), 471-487.

19.   Haugh, M. (2007). The discursive challenge to politeness research: An interactional alternative.

20.   Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Sanchez, V., & Perez-Meana, H. J. a. p. a. (2018). A web scraping methodology for bypassing twitter API restrictions.

21.   Jiang, X., Hu, Y., & Li, H. (2009). *A ranking approach to keyphrase extraction.* Paper presented at the Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.

22.   Kelly, B., & Nixon, W. J. O. R. (2013). SEO analysis of institutional repositories: what's the back story? , *505*.

23.   Kim, H., Baik, S. Y., Yang, S. J., Kim, T. M., Lee, S. H., Cho, J. H., . . . Toxicology. (2019). Clinical experiences and case review of angiotensin II receptor blocker-related angioedema in Korea. *124*(1), 115-122.

24.   Kothari, S., Mizik, N., & Roychowdhury, S. (2012). *Managing for the moment: The role of real activity versus accruals earnings management in SEO valuation*. Retrieved from

25.   Landim, C., Mariani, M., Seo, I. J. A. f. r. m., & analysis. (2019). Dirichlet's and Thomson's principles for non-selfadjoint elliptic operators with application to non-reversible metastable diffusion processes. *231*(2), 887-938.

26.   Madni, H. A., Anwar, Z., & Shah, M. A. (2017). *Data mining techniques and applications—a decade review.* Paper presented at the 2017 23rd International Conference on Automation and Computing (ICAC).

27.   Salager-Meyer, F. J. T.-I. j. f. t. s. o. d. (1990). Discoursal flaws in medical English abstracts: A genre analysis per research-and text-type. *10*(4), 365-384.

28.   Shi, J., Cao, Y., & Zhao, X.-j. (2010). *Research on SEO strategies of university journal websites.* Paper presented at the The 2nd International Conference on Information Science and Engineering.

29.   Vargiu, E., & Urru, M. J. A. I. R. (2013). Exploiting web scraping in a collaborative filtering-based approach to web advertising. *2*(1), 44-54.

30.   Vaughan, L., Shaw, D. J. J. o. t. A. S. f. I. S., & Technology. (2005). Web citation data for impact assessment: A comparison of four science disciplines. *56*(10), 1075-1087.

31.   Wilson, A. M., Thabane, L., & Holbrook, A. J. B. j. o. c. p. (2004). Application of data mining techniques in pharmacovigilance. *57*(2), 127-134.

32.   Zhao, C., Lu, J., & Duan, F. (2009). *Application and Research of SEO in the Development of Web2. 0 Site.* Paper presented at the 2009 Second International Symposium on Knowledge Acquisition and Modeling.