JADAVPUR UNIVERSITY

MASTER DEGREE THESIS

COMPUTATIONAL ANALYSIS OF WEB PAGE METRICES FOR SEARCH ENGINE OPTIMIZATION

A thesis is submitted in partial fulfillment of the requirements for the degree of
Master of Technology in Computer Technology
in the
Department of Computer Science and Engineering

By

Arpan Mondal

University Roll Number: **001910504006**
Examination Roll Number: **M6TCT22008**
Registration Number: **149841** of **2019-20**

Under the Guidance of
Prof. (Dr.) Subhadip Basu
Department of Computer Science and Engineering
Jadavpur University
Kolkata-700032

2021-2022

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Certificate of Recommendation

This is to certify that the dissertation entitled **"Computational Analysis of Web Page Matrices for Search Engine Optimization"** has been carried out by **Arpan Mondal** (University Roll No: **001910504006**, Examination Roll No: **M6TCT22008,** University Regn No: 149841 of 2019-20), under the guidance and supervision of **Prof. (Dr.) Subhadip Basu,** Department of Computer Science and Technology, Jadavpur University, Kolkata, is being presented for partial fulfilment of the Degree of Master of Computer Technology during the academic year 2020-202 1. The research results presented in the thesis have not been included in any other paper submitted for the award of any degree in any other university or institute.

Prof. (Dr.) Subhadip Basu (Thesis Supervisor)
Department of Computer Science and Engineering
Jadavpur University, Ko1kata-32

Countersigned

Prof. (Dr.) Anupam Sinha
Head of Department,
Computer Science and Engineering,
Jadavpur University, Kolkata-32.

Prof. Chandan Mazumdar
Dean,
Faculty of Engineering and Technology
Jadavpur University, Kolkata-32.

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Certificate of Approval
(Only in case the thesis is approved)

This is to certify that the dissertation entitled **"Computational Analysis of Web Page Matrices for Search Engine Optimization"** is a bonafide record of work carried out by **Arpan Mondal** (University Roll No: **001910504006**, Examination Roll No: **M6TCT22008,** University Regn No:**149841** of **2019-20**) in partial fulfillment of the requirement for the Degree of Master of Computer Technology in the Department of Computer Science and Engineering, Jadavpur University duringthe period of August 2021 to July 2022. It is understood that by this approval the undersigned does not necessarily endorse or approve any statement made,opinion expressed, or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

Signature of Examiner 1
Date:

Signature of Examiner 2
Date:

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Declaration of Originality and Compliance of Academic Ethics

I hereby declare that this thesis entitled **"Computational Analysis of Web Page Matrices for Search Engine Optimization"** contains a literature survey and original research work by the undersigned candidate, as part of his degree of Master of Computer Technology.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name (Block Letters): **ARPAN MONDAL**

Roll Number: **001910504006**
Examination Roll Number: **M6TCT22008**
University Registration Number: **149841 of 2019-20**

Thesis Title: **COMPUTATIONAL ANALYSIS OF WEB PAGE METRICES FOR SEARCH ENGINE OPTIMIZATION**

Signature with Date:

Acknowledgment

I would like to express my deepest gratitude and sincere thanks to my respected mentor and teacher Prof. (Dr.) Subhadip Basu, Department of Computer Science and Engineering, Jadavpur University for his exclusive guidance and undivided support in completing and presenting the thesis successfully. I am indebted to him for the constant encouragement and continuous knowledge sharing that he has given to me. The above words are only a token of my deep respect towards him for all that he has done to make my thesis its present shape.

I would also like to thank Mr. Kinshuk Das Sir for the valuable knowledge-sharing sessions and guidance that he has provided during the complete duration of this thesis, without whom completing this thesis wouldn't have been possible. I am grateful to him for everything he has done during the work.

I would also like to thank my friends Sharmistha Ghosh, Ipsita Das and Sneha Poddar for constantly supporting and guiding me through the research work.

I would also like to thank my colleagues from CTS and Citicrop India pvt ltd for support and motivate me throughout the time.

Finally, I express my sense of gratitude and thanks to my parents for unconditional love and support, and without whom the whole work wouldn't have been possible. Last but not least I am grateful to all my friends, teachers, and mentors who have helped me in the work and provided me with key insights.

Regards,

ARPAN MONDAL

University Roll Number: **001910504006**
Examination Roll Number: **M6TCT22008**
University Registration Number: **149841 of 2019- 20**
Department of Computer Science and Engineering
Jadavpur University

Signature with Date:

Abstract

The relevance of a web page (based on search keywords) is a naturally biased topic that is entirely dependent on the knowledge, interests, and attitude of the reader. Any web page is reliant on these elements. The fast growth of the internet is one element that makes it challenging for search engines to provide relevant results to consumers in a timely manner. Search engines employ classified directories to store webpages, and some search engines even rely on human expertise in this process. The majority of web pages classify their content using automated algorithms.

This paper proposes a novel technique to analyzing SEO attributes of top 5 sites in google (based on keywords search on google) and making predictions about how to use suitable SEO attributes to improve site ranking on google search using machine learning algorithms based on experts' experience. The findings of the experiments suggest that machine learning may be used to anticipate the degree of web page adaptation to SEO advice.

The proposed approach has practical value in that it provides the foundation for developing software agents and expert systems that can automatically detect web pages, or parts of web pages, that need to be improved in order to comply with SEO guidelines and, as a result, gain higher search engine rankings. The findings of this study also contribute to the topic of determining the best values for ranking criteria used by search engines to rank web pages. Experiments in this study reveal that the page title, meta description, H1 tag (header), and body text are essential aspects to consider when creating a web page, which is consistent with earlier studies. Another outcome of this study is a new data set based on machine learning web page prediction that can be used in future studies.

Contents

List of Figures

Chapter 1

Introduction

Search engines (SEs) are a popular way to access content on the Internet. Users can utilize search engines like Google, Baidu, and Yandex to discover the most relevant information for their query. SEs must sift through a massive amount of data to accomplish this. SEs crawl websites' text content via links and store it in their database ('index ') for later analysis. SEs are thought to use powerful computational techniques to evaluate the relevance of web pages for search queries, while the details are uncertain.

A search engine (SE) is a programme that allows users to find information on the Internet. Users type keywords into search engines to identify Web sites that have the information they're looking for. Search engine optimization (SEO) is a set of tactics that help your website rank higher in organic (or "natural") search results, making it more visible to individuals searching for your product or service on the internet. [1] Search Engine Optimization / Search Engine Marketing is the process of making a site and its content highly relevant for both search engines and searchers, and it is a type of online marketing [2].

A site's top ranking for relevant terms and phrases is aided by effective search marketing. Search Engine Optimization (SEO) is a subset of Search Engine Marketing (SEM), a term that encompasses all search marketing tactics. Both organic and paid search are included in SEM. You can pay to have your website listed on a search engine so that it appears when someone types in a specific keyword or phrase [3]. On the search engine, both organic and paid listings appear, but they are presented in various places on the page [4].

The technique of altering the appearance of a website or a web page in a search engine's "natural" or unpaid ("organic") search results is known as search engine optimization (SEO). In general, the earlier (or higher placed on the search results page) and more frequently a site shows in the search results list, the more visitors the search engine's users will see it. Image search, local search, video search, academic search, news search, and industry-specific search engines are all possible targets for SEO [1]. SEO evaluates how search engines work, what people search for, and the actual search phrases or keywords typed into search engines as an Internet marketing technique.

1.1 Search Engine

A web search engine is a software system that searches the World Wide Web for information. The search results are typically displayed in a list of results known as search engine results pages (SERPs). The data may consist of a combination of web pages, images, and other types of files. Some search engines mine data from databases or open directories. In contrast to web directories, which are only maintained by human editors, search engines maintain real-time information by running an algorithm on a web crawler [6].

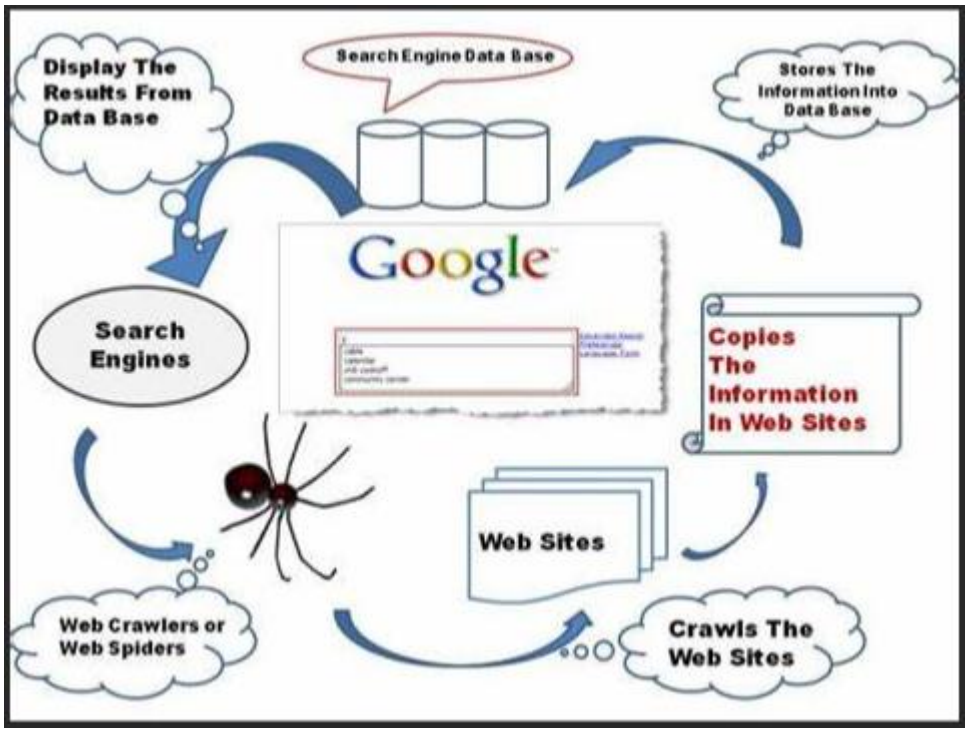


Figure 1: Structure of Search Engine

Web search engines operate by storing information about a large number of web pages, which they retrieve from the HTML markup of the pages. A Web crawler (also known as a spider) — an automated Web crawler that follows every link on the site — retrieves these pages. The site owner can use robots.txt [1] to exclude specific pages.

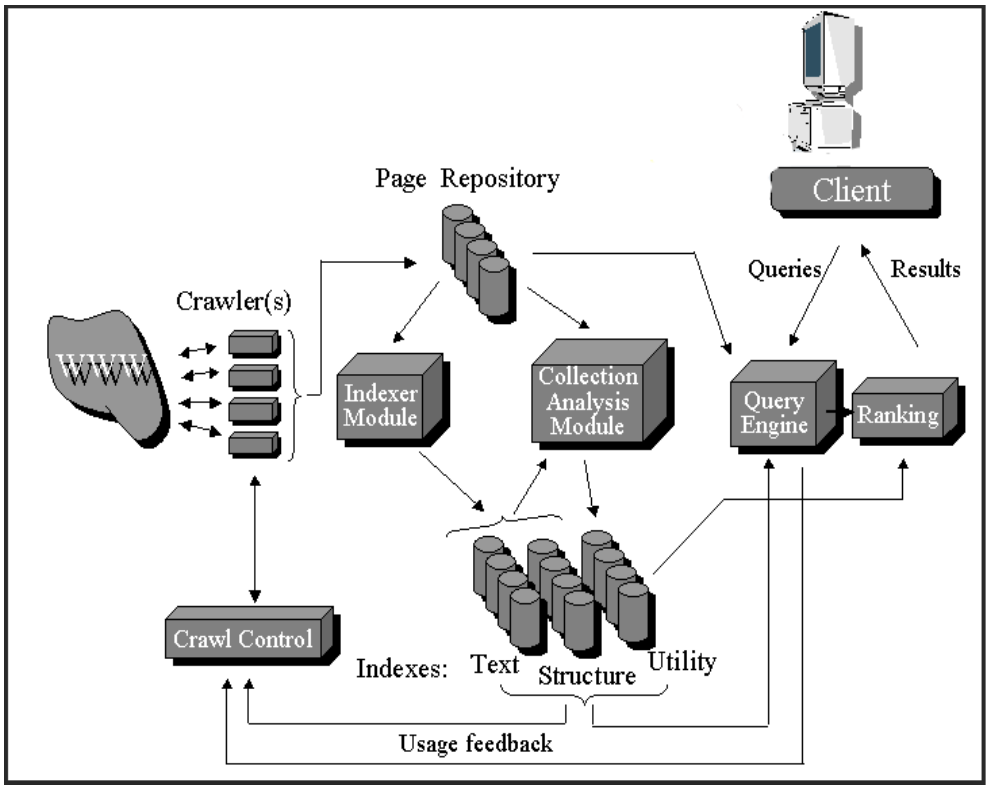


Figure 2: Search Engine Architecture

With the help of internet resource, the web crawlers find fetches the web pages and those web pages are indexed and are sorted by words in the database which is further process towards the Google database for query processor which is further carried towards search engine for response [5].

1.2 Why is Ranking so important?

For obvious reasons, the primary goal of any website is to be ranked first. Search engines are a strong tool that customers use to locate information online. They are the most efficient and important way for them to do so. When searching, the majority of individuals will begin at the top of the page and scroll down until they find what they are looking for. Less experienced internet users may blindly believe the search engine's recommendations and will only visit a few websites to obtain the information or service they require, therefore being at the top is crucial.

According to research, people are 40% more likely to click on your site if it is rated first, 30% more likely if it is ranked second, and 24% more likely if it is ranked third [3]. It's simple for more experienced surfers to type in keywords or a business name and see what comes up, but if you're not at the top, or even on the first page of results, you're missing out on traffic, site visits, and most importantly, prospective business.

1.3 Search Engine Optimization

The functionality and effectiveness of web pages are also affected by their ranking in search engine results, which is determined by a variety of factors [7]. These factors are well known, but their impact on rank has not been thoroughly documented. Webmasters and web authors known as search engine optimization experts (SEO experts) can easily estimate whether a specific web page has been optimized for specific keywords and follows the SEO guidelines published by the popular search engines based on their experience and knowledge. SEO guidelines are recommendations on how to construct web pages so that search engine algorithms can understand better and rank them [8].

Figure 3: Various Search Engines

Much of the previous SEO research has concentrated on the significance of links in search rankings. This is due to the importance of link quantity and quality in PageRank, Google's core algorithm [9]. However, considerably fewer studies investigate the impact of content and textual features on search rankings in a way that site creators can evaluate. We address that gap in this study by asking, "How do different content features predict search ranking of online content websites?"

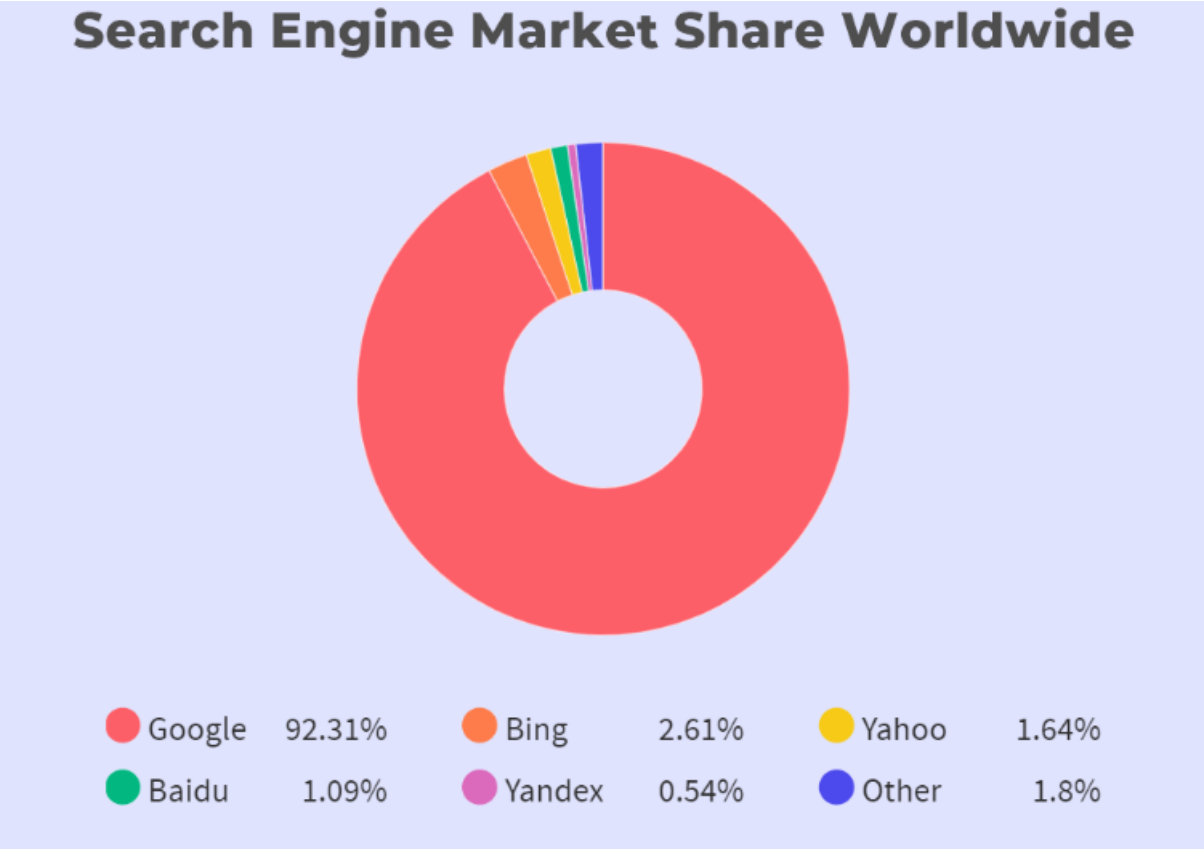


Figure 4: Search Engine Market Share Worldwide

We use predictive machine learning to answer this question. To understand which factors, boost a website's position in search results, we use a dataset of search queries and their search engine rankings. The data contains information on the content and link profiles of several websites. We can use this information to aim top in point what features contribute to a page's position in the search results for a relevant query.

Chapter 2

SEO AND SEARCH ENGINES

Although search engines are unquestionably important in SEO, many businesses are unaware of how they operate. Websites host a variety of HTML documents, each with its own Uniform Resource Locator (URL) (URL). A search engine facilitates Web searching by creating an index, which is invisible to the user, and responding to queries, which requires the user's active participation.

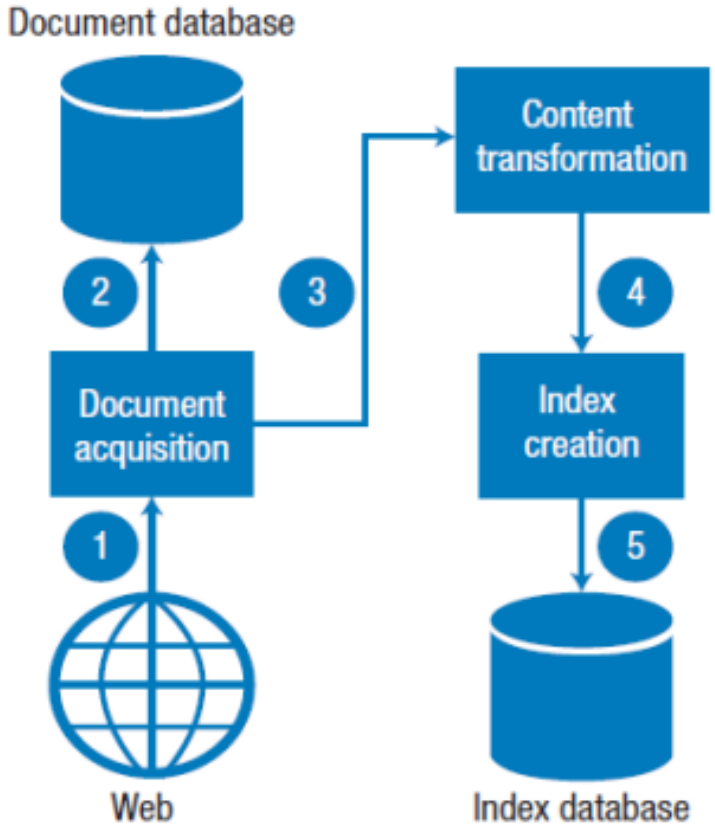


Figure 5: Indexing. Indexing involves fetching HTML documents, storing them in their original form, transforming the documents by processes such as stop word removal and stemming, and generating indexes and storing them in a database

Figure 5 depicts a conceptual diagram of web page (or document) indexing. The numbers in the circles correspond to the following steps:

A web crawler, also known as a robot or bot, searches the Internet for HTML pages.
The web crawler saves these pages in their original form in the document Database of the search engine.

The pages are subjected to transformations such as HTML tag and stop-word removal, as well as stemming. The search engine's transformation extracts significant textual content and link information for indexing.
Indexes are created by the search engine by generating direct and surrogate page representations, such as single words or phrases and their position on the page. It also keeps track of incoming and outgoing links and generates a snippet.
The search engine stores indexes in its index [10] database.

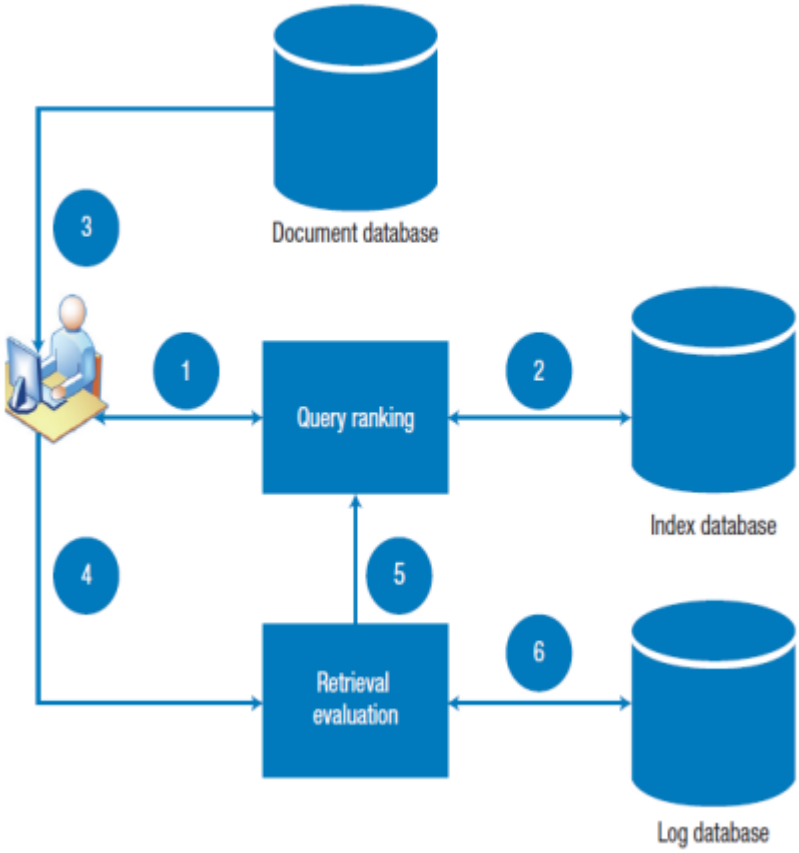


Figure 6: Querying and document retrieval. From a user query, an algorithm generates a ranked list of relevant documents, from which the user browses retrieved documents by clicking on the corresponding links. Queries can be refined and re-executed on the basis

The querying and document-retrieval process is depicted schematically in Figure 6. The numbers in the circles correspond to the following steps:

1. The user uses the browser of a search engine to enter a search query, which is typically a single keyword or short phrase [10]. The search engine transforms the user's query in canonical representation, as it did in step 3 of the indexing process.

2. The query-ranking algorithm of the search engine generates a ranked list of URLs for documents that it deems relevant based on the index database and contextual information in the user query. The search engine then displays the snippets associated with the ranked URLs to the user in SERPs.

3. The user browses the snippets and clicks on specific ones to retrieve the full documents in their original form.

4. The retrieval evaluation component of the search engine assists the user in further refining the search based on feedback about the document's relevance: the user explicitly indicates relevance (direct feedback) or clicks on relevant links (indirect feedback).

5. Using the relevance feedback, the search engine may reformulate and re-execute the user query. This process is repeated until the user is satisfied with the search results or the query session is terminated.

6. The search engine stores Meta information in the log database, such as user queries, relevance feedback, and clicked snippets, which it uses to improve search performance.

Chapter 3

Literature Survey

3.1 Search Engine Optimization

Zhang and Dimitroff [11] investigated the impact of metadata implementation on webpage visibility in search engine results. They modified metadata of 46 web pages and submitted the modified versions to 19 search engines (among them Google); over several weeks, the rankings were improved in eight of the search engines. Zhang and Dimitroff [11] found that pages with metadata had higher visibility than pages without metadata, especially when the metadata was mentioned also in the page's text content. Regarding content characteristics, Zhang and Dimitroff [12] found that using duplicate keywords in the title increases the ranking but only until three repetitions, after which there is a decrease in visibility. The same effect was found for the body text, except without a decrease from repetitive use of the keyword. Moreover, using keywords in both title and body text resulted in better performance than using the keywords in just either of the two, while changes in font color, case, size, or use of plural and adjectives did not affect ranking [12].

Evans [13] analyzed Google rankings of 50 optimized and 50 non-optimized web pages. The researcher found the following factors influential for higher rankings: a) PageRank score, b) the number of inbound links, c) age of the domain name, and d) listings in Yahoo! and DMoz directories. Malaga [14] conducted an experimental study to increase the search rankings of an ecommerce site by using on-site optimization and link building tactics, finding that links from reputable websites had a major impact. Wang et al. [15] collected data from 118 websites to measure the effect to franking factors, including a number of inbound links, title length, and keyword density. They found link popularity as the most important criteria. Other recommendations included limiting website title to 80 characters, page size to smaller than 150 kilobytes, the hierarchical order in a directory listing less than 4, and keyword density between 2– 8% [15].

Gasparotto [16] suggests that higher website rank is correlated with higher site visit numbers, resulting in an effect where big websites are able to maintain their positions. Lee et al. [18] conducted a case study to analyze the SEO

techniques applied to the LG Science Land content, finding the following factors influential: (a) simplified URL structure, (b) internal redirect in the case of page removal, (c) XML sitemap to help search engines index the site, (d) descriptive title and meta-tags, (e) use of canonical URLs, and (f) removal of expired links and content [18]. Zhang and Cabage [20] compared the effect of link building and social sharing on search rankings. They analyzed three content-rich websites with similar content, site and page structure, the volume of traffic, and search rankings. SEO efforts, including content creation, link building, and social sharing were then applied to treatment websites while the control sites were left as they were. The findings showed that links had the strongest impact on ranking over 18 months. While social sharing had a rapid impact on traffic, the increase was proved having only a temporary effect on the search ranking [20].

SEO Factor	Impact on ranking		
	Positive	**Negative**	**Neutral**
Meta tags	[7][15]		[14]
Age of domain	[9]		
Internal links	[15]		
External links	[9][19][14][16]		
Number of pages			[9]
Use of keywords in body text	[14][8]		
Use of directories	[9][14]		
Page file size	[14]		
Social sharing			[16]
Website traffic	[9]		
* When used excessively			

Figure 7: SEO Factor impact on Ranking

Overall, there has been surprisingly little empirical work on the effect of search-engine optimization techniques despite the impact of search engines on the revenue of companies. The earliest work [7, 8] emphasizes the use of meta-tags, after which the focus shifts on content, and particularly giving more exact prescriptions on the length and densities of content elements in relation to keywords [15]. The changes in the emphasis of the earlier studies are descriptive of the ever-changing nature of search-engine algorithms. Links have remained as essential part of SEO research, mostly due to Google's PageRank algorithm emphasizing reputable inbound links [9].

3.2 Techniques of SEO

For web page rankings, search engines employ algorithms that take into account a variety of factors. The majority of these factors are well-known, but the SEO guidelines do not specify their role or impact on ranking. Search engines publish guidelines on how to use these factors, but they do not reveal the golden rule or formula that will get a website to the top of a search results listing. On-page factors and off-page factors are the two types of ranking factors.

3.2.1 On-Page Factors:

On-page factors are web page content characteristics that the webmaster has complete control over. Web page text, links, images, tables, navigations, URLs, file names, and HTML code are examples. The following on-page factors are highlighted in search engine guidelines [8].

- Text quality (including information quality),
- Clear navigation,
- Page title (HTML "title" tag),
- Meta-description (HTML "meta description" tag),
- Using H tag for marking titles (H1, H2, H3, etc.),
- ALT attribute on images (short image description),
- Anchor text of links,
- URL (address) of the page (including the domain name),
- Web page loading speed, and
- HTML code-to-content ratio (must be in favour of the content).

This is not an exhaustive list. A set of on-page factors includes many more factors that have less influence (like the domain age, for example). All of these factors have one thing in common: they can be controlled, unlike off-page factors, which cannot be controlled to the same extent.

3.2.2 Off-Page Factors:

Off-page factors, or factors that are unrelated to the content of a web page and rely on external influences, are important factors that influence today's search engine algorithms. These are the factors that are largely beyond the web page author's control. Incoming and outgoing links, comments [17], as well as their quality and recommendations from social media sites [18], digital marketing [19] are among the most important off-page factors. PageRank [20] and HITS [21] were two of the first algorithms to use links to rank web pages. They regarded links as votes, assuming that web pages with more links were more important and should thus appear higher in search results.

Off-page factors will not be discussed in detail since they were not the subject of the research in this paper.

3.3.3 Keywords:

The SEO process always starts with keywords, which are one or more words that Internet users type into search engines when looking for information. They are search queries that can also be formulated as questions that can be answered by today's search engines. Nonetheless, using keywords rather than questions is still a more common method of searching for information on search engines [22]. When searching on search engines, most users use queries with two to three words, which is considered the best approach.

For a web page to appear in organic search engine results' pages, it needs to contain the keywords in both on-page and off-page factors. The basic idea behind SEO is to find keywords that will push the web page high in search engine result pages.

Keywords [24] are important in various fields of study. They are used to find spam webpages, malicious content, fake news, and to trick search engines' algorithms by employing unethical techniques called "black hat SEO". To classify web pages based on text topic, or to suggest keywords for SEO process, keyword suggestion systems were developed. In this study, keywords were not extracted from page body, but from catalogue category

Chapter 4

Analysis of Various Metrices for SEO

The main questions of this research are to analyse any publicly available website and investigate the different SEO metrices and their related studies. Here we have developed an application that will analyse the different SEO metrices and as well as compare those results with top 5 publicly available sites based on search keywords. The comparison will give you a comparative study on important of SEO attributes and based on the results you can actually update your website to have a better ranking in google.

Figure 8: Application Home Page

4.1.2 Analysis of Images

4.1.2.1 Introduction:

We have taken all of the images from a website and calculated the number of images in the website, the maximum size of all the images, the minimum size of all the images, the sum of all the images, the mean of all the images, and the standard deviation of all the images using this technique. In addition, any photos with length issues or contradictions have been removed. Finally, it's time to assess the overall performance of the images.

4.1.2.2 Dependencies:

If the number of images is too much in a website or size of images are big then it will take more time to load and that leads the high load time of a website. It will cause low ranking of a website. After analysis it will show overall performance that's decide performance of web page.

4.1.2.3 Process:

Here sd is Standard value of image in web. After made long web search we make a decision that sd value must be within 100KB

y=(VAL-MIN)/(C1-MIN)

After adding all the value of x and y then divided by total number of images will give overall performance.

MEAN = Total sum of images / Number of Images

$$\text{STANDARD DEVIATION} = \sqrt{\frac{\sum(x-mean)^2}{2N}}$$

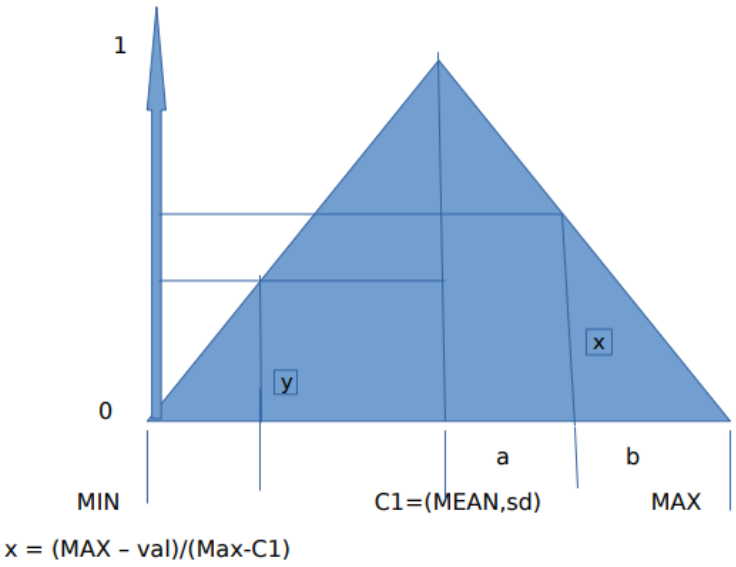


Figure 9: Calculation of Image SD,mean

4.1.2.4 Experiment Result:

Input: This is a result of experiment that clearly describe the image analysis of that particular websites.

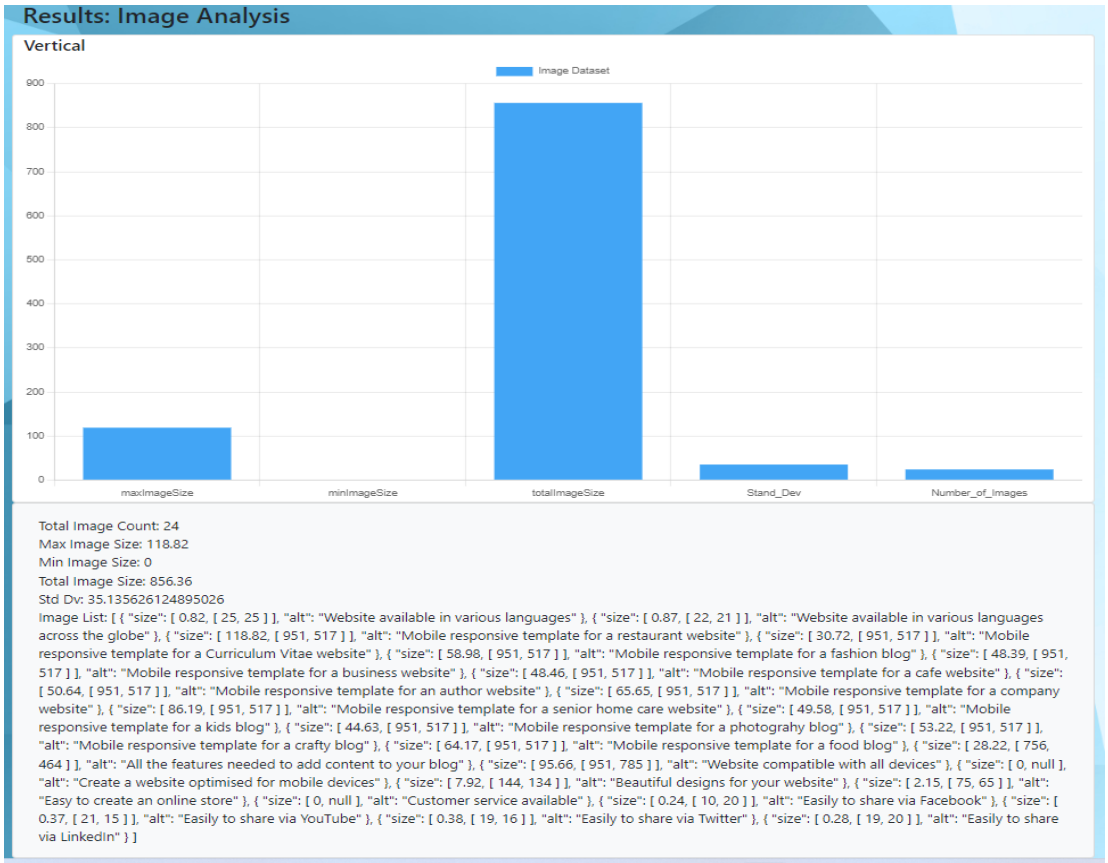


Figure 10: Image analysis from Application

4.1.3 Analysis of Paragraphs

4.1.3.1 Introduction:

We have taken all of the paragraphs from a website and calculated the number of paragraphs in the website, the maximum size of all the paragraph, the minimum size of all the paragraph, the sum of all the paragraph, the mean of all the paragraph, and the standard deviation of all the paragraph using this technique. Finally, it's time to assess the overall performance of the paragraph.

4.1.3.2 Dependencies:

After analysis it will show overall performance that's decide performance of web page. How paragraph should be maintained in a web page. Number of relative keywords in paragraph will maintain the good ranking of a website.

4.1.3.3 Process:

y=(VAL-MIN)/(C1-MIN)

After adding all the value of x and y then divided by total number of paragraph will give overall performance.

MEAN = Total sum of paragraph / Number of paragraph

$$\text{STANDARD DEVIATION} = \sqrt{\frac{\sum(x-mean)^2}{2N}}$$

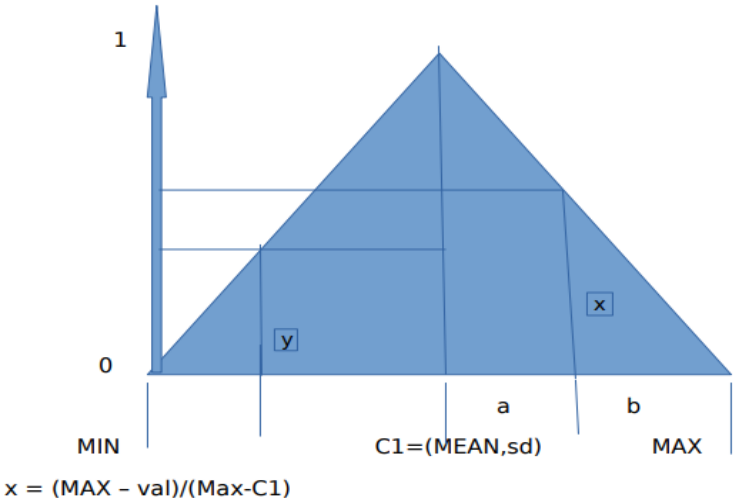


Figure 11: Calculation of Paragragh Analysis

4.1.3.4 Experiment Result:

Input: This is a result of experiment that clearly describe the paragraph analysis of that particular websites.

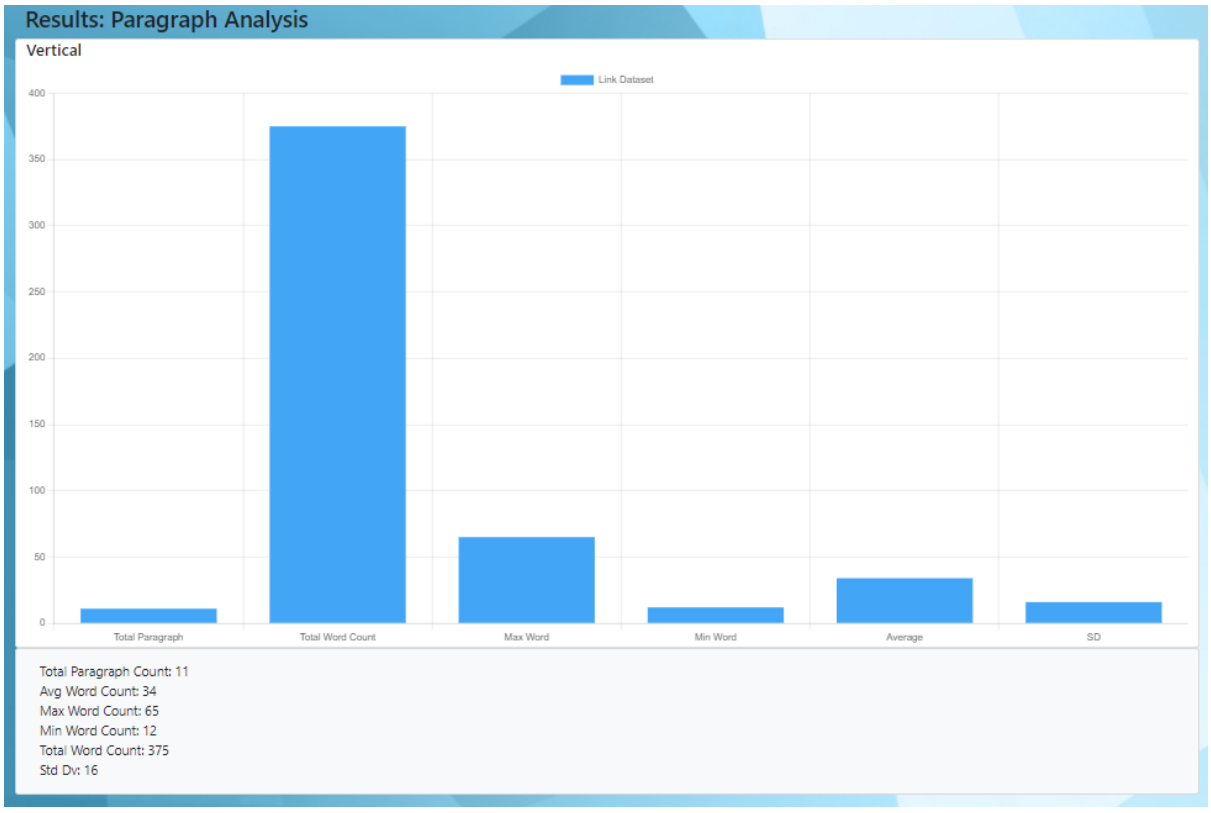


Figure 12: Paragraph analysis result from Application

4.1.4 Analysis of Internal and External Link

4.1.4.1 Introduction:

One of the most important aspects of a website that most website owners worry about is the number of links they have on their web pages, both internal and external links. That's why we made an analysis of total number of links in a web page, Total Number of Internal Link and Total Number of External Link. The backlinks your website, the higher is the website authority.

4.1.4.2 Dependencies:

Because the more and quality links a website has, the higher it will be ranked by search engines like Google.

4.1.4.3 Process:

Internal links: are link pages that can be found on a website. They are commonly linked through the website menu or by interlinks.

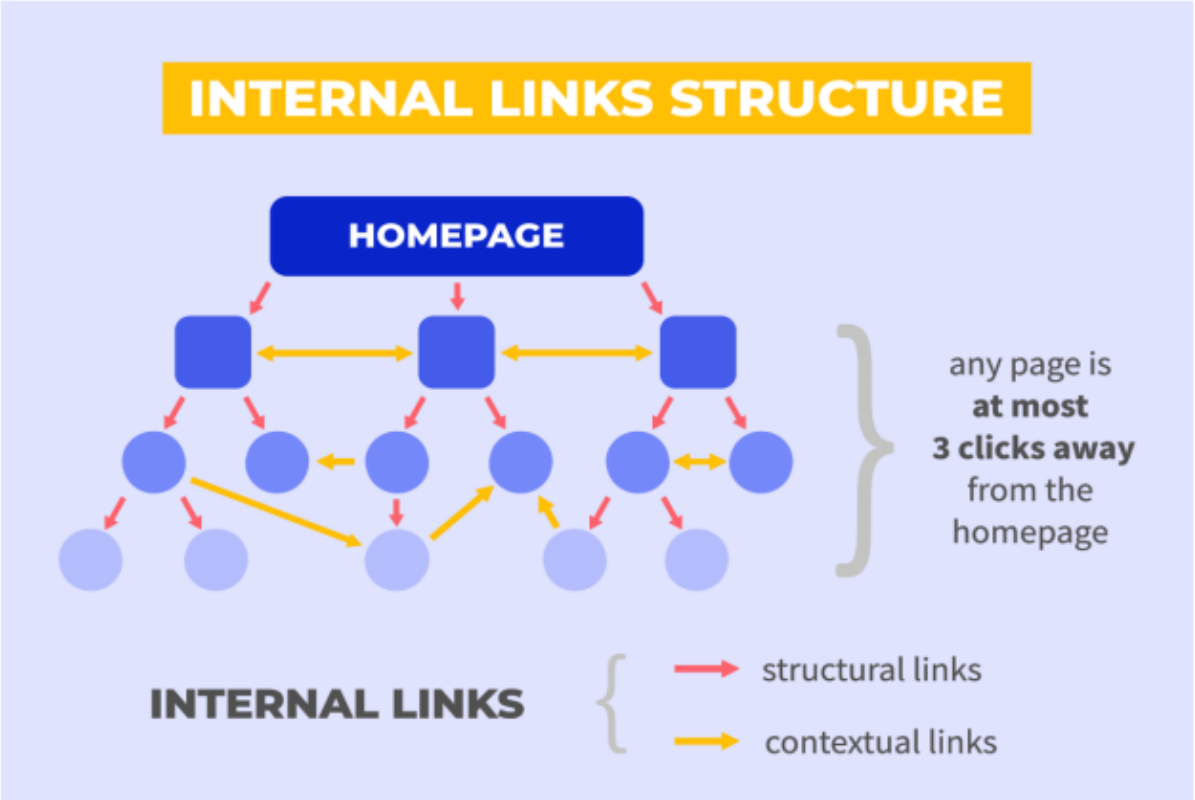


Figure 13: Internal Links Structure

External links: are the links to other websites or also known as partner websites. The higher the number of external links that a website has the better for search engine optimization. But all links should be with relevant and reliable websites and not spam sites.

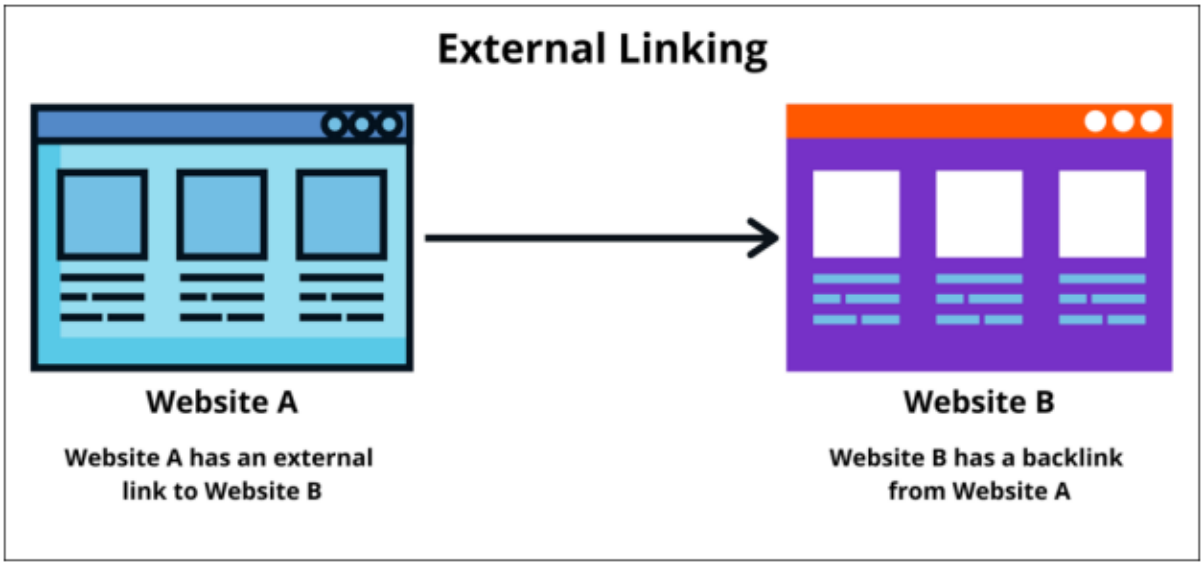


Figure 14: External Links Structure

4.1.4.4 Experiment Results:

Figure 15: External Link and Internal Link analysis for an website

4.1.5 Analysis of No Follow and Do Follow Link

4.1.5.1 Introduction:

One of the most important aspects of a website that most website owners worry about is the number of links they have on their web pages, both do follow and no follow links. That's why I made an analysis of total number of links in a web page, Total Number of Do Follow Link and Total Number of No Follow Link.

4.1.5.2 Dependencies:

Because the more and quality links a website has, the higher it will be ranked by search engines like Google.

4.1.5.3 Process:

No follow link - are typically shown in red as they indicate threat. The extent of threat that they have to a website is still debatable. Still, we must avoid no follow links. Popular social media sites like Facebook and Twitter are automatically No-follow.

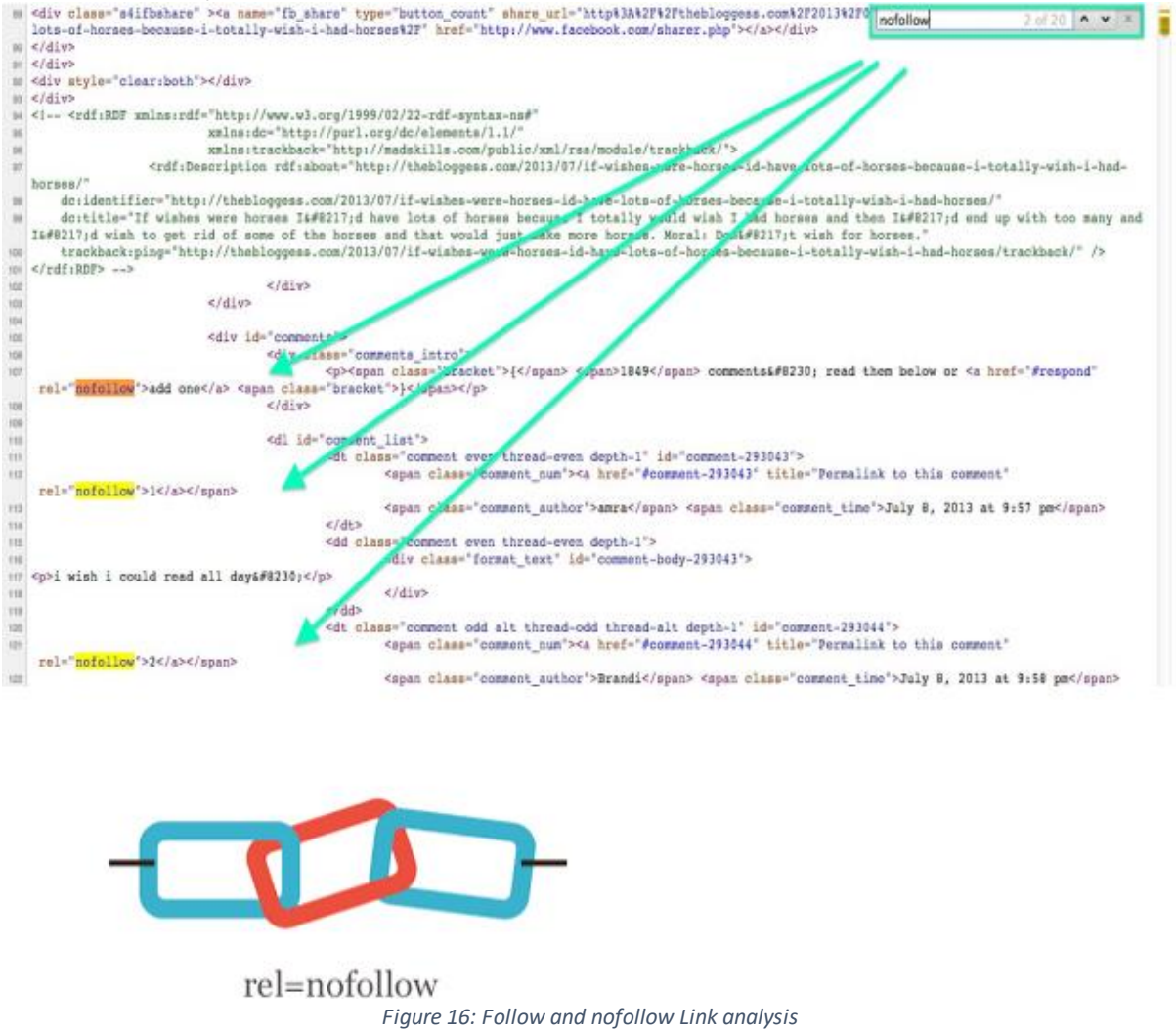


Figure 16: Follow and nofollow Link analysis

Do follow link – are links that allow Google and other search engines to follow them and reach your website. This is link is beneficial for the website because it can help us to get a good page ranking on search engines.

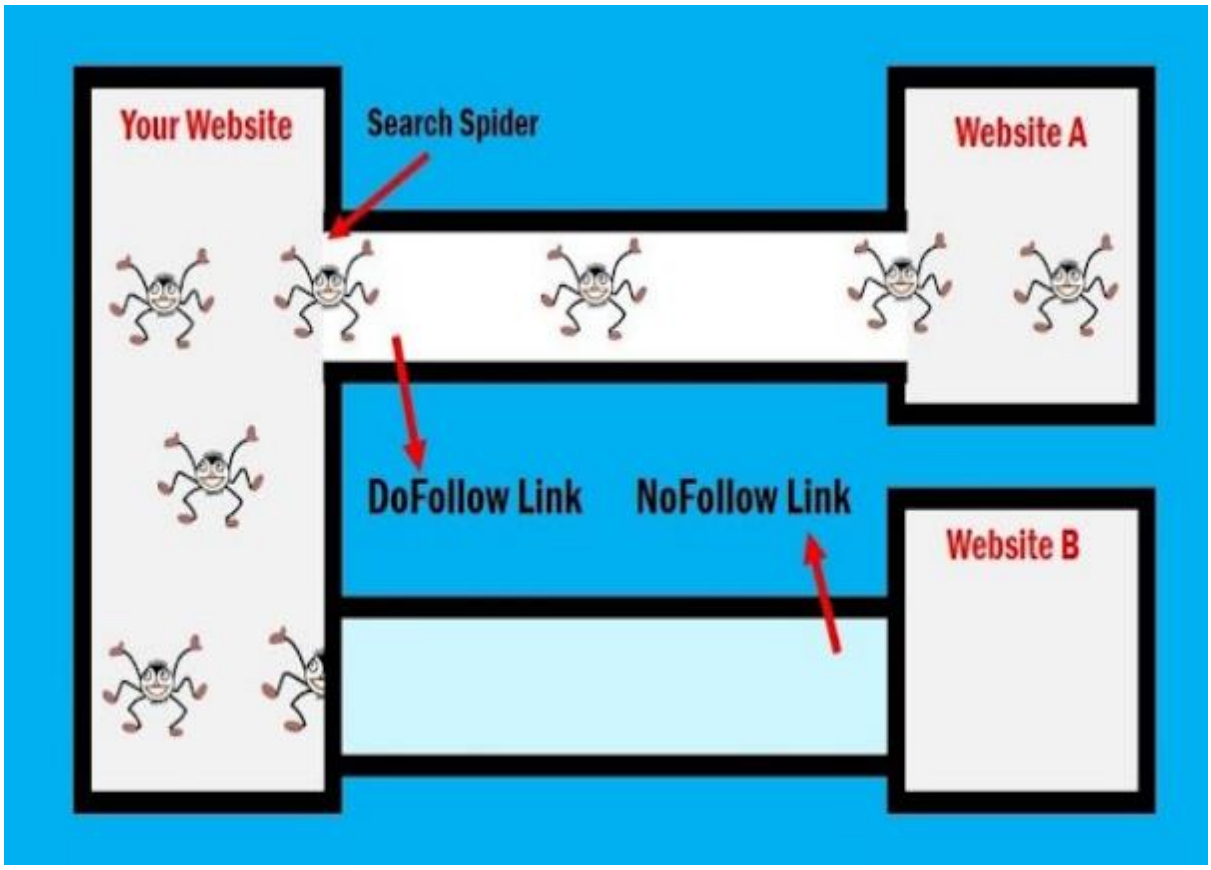


Figure 17: DoFollow vs NoFollow

4.1.5.4 Experiment Result:

Figure 18: Follow and Nofollow analysis

4.1.6 Meta Description search and Analysis

4.1.6.1 Introduction:

A meta description is the small blurb that appears underneath your website on the SERP that includes information about your page.

4.1.6.2 Dependency:

It is designed to provide users with a brief summary of the content on your page so that users know if the page will answer their question. There is a standard value of length of description that is under 155-160 characters.

4.1.6.3 Process:

It is just picking up all meta tag where name is present and name is description.

4.1.6.4 Experiment Result:

Figure 19: Meta Description Analysis

4.1.7 Broken Link Analysis

4.1.7.1 Introduction:

When we click on a link that is supposed to take you to a particular page, yet instead it takes you to another page that shows a 404-error message - this is called a broken link.

Broken Link is a crucial factor in SEO. Broken links make the user experience very unpleasant, and can damage the reputation of your website. It makes sense that having several broken links on a website is often referred to as “link rot” - because it is as bad as the phrase sounds. It can effect in Page Ranking.

4.1.7.2 Process:

Listed below are some of the most common causes for a broken link:

- The website is temporarily or permanently unavailable
- The web page has been deleted
- The web page was modified or changed
- The web page was blocked by firewall or other similar software

So, the process of finding broken is simple. Where I find out all the links of a website and then checked their status code if its 404 then it’s a broken Link.

4.1.7.3 Experiment Result:

URL	https://www.bloomnation.com/
Broken Links Are	
https://www.bloomnation.com/blog/rss/index/store_id/1/	

Figure 20: Broken Link Analysis

4.1.8 Website Summarizer

4.1.8.1 Introduction:

In the last two decades, text summarization has played an essential role in search engine optimization (SEO). There are, indeed, a lot of different marketing techniques that require a summary of the content, and that can improve ranking in search engines. *Meta descriptions* are probably among the most notable examples

These text snippets provide search engines with a brief description of the page content and are still an important ranking factor and one of the most common use cases for text summarization.

Thanks to the latest NLP technologies, SEO specialists can finally summarize the content of entire webpages using algorithms that craft easy-to-read summaries.

Summaries help create and structure the metadata that describes web pages. Text summarization also comes in handy when we want to add descriptive text to category pages for magazines and eCommerce websites or when we need

to prepare the copy for promoting our latest article to Facebook or Twitter. Much like search engines use meta descriptions, social networks rely on their meta descriptors like the Facebook Open Graph meta tag (a.k.a. OG tag) to distribute content to their users. Facebook for instance, uses the summary provided in OG tags to create the card that promotes a blog post on mobile and desktop devices.

4.1.8.2 Process:

There are many different text summarization approaches, and they vary depending on the number of input documents, purpose, and output. But, they all fall into one of two categories: **extractive** or **abstractive**.

Extractive summarization algorithms identify essential sections of a text and generate verbatim to produce a subset of the sentences from the original input.

Abstractive methodologies summarize texts differently, using deep neural networks to interpret, examine, and generate new content (summary), including essential concepts from the source.

4.1.8.3 Experimental Results:

Figure 21: Website Summarizer Analysis

4.1.9 Header Analysis

4.1.9.1 Introduction:

In the early 2000s, heading elements (H1, H2, H3) were actual ranking factors. It was mandatory to add your keywords in the headings if you wanted to rank. But that's not been the case for many years. Yet, it is a common SEO practice to worry about keywords in the heading tags.

The word "rote" means doing something mechanically and out of habit, without thinking about it. Adding keywords to headings tags has become a rote SEO practice. It's done not because it's useful but because it's a habit. Take a look at the top ranked sites for most any query and it's highly likely you won't see sites seeding heading tags with keywords.

Heading tags have made the top ten lists of ranking factors for several decades. But if you look at the search engine results pages (SERPs) you'll see that's not the case. Anyone who argues otherwise is denying what exists in front of their eyes.

The proper use of heading tags has changed. The proper use is to help search engines understand what the content is about. That's it.

4.1.9.2 Dependency:

Both the title and meta description should be included in the section of every page on your website. When it comes to your pages' body text, one of the best things you can do to improve your overall SEO value is to include and heading tags containing your target keyword phrases in your content. Not only do the search engines place added weight on the words found in these particular tags, the visual relief they provide to your website's readers will help them to move more efficiently through your site's content.

4.1.9.3 Experimental Results:

Results: Header Analysis

H1

Text	Size
How do I solve NameError: name 'threading' is not defined in python 3.3	71
Subscribe to RSS	16

H2

Text	Size
1 Answer 1	47
Your Answer	11
Not the answer you're looking for? Browse other questions tagged python python-3.x or ask your own question.	109

H3

Text	Size
current community	17
your communities	16
more stack exchange communities	31
Sign up or log in	17
Post as a guest	15
Post as a guest	15

H4

Text	Size
Related	7
Hot Network Questions	21

H5

Text	Size
Stack Overflow	14
Products	8
Company	7
Stack Exchange Network	22

Figure 22: Header Tag Analysis

4.1.10 Canonical Tag Checking and Analysis

4.1.10.1 Introduction:

A canonical tag (“rel=canonical") is a way of telling search engines that a specific URL represents the master copy of a page

4.1.10.2 Dependency:

Using the canonical tag prevents problems caused by identical or "duplicate" content appearing on multiple URLs.

4.1.10.3 Process:

It’s just pick up all the link tag where rel is canonical.

4.1.10.4 Experiment Result:

Figure 23Canonical tag Analysis

4.1.11 robots.txt File Checking

4.1.11.1 Introduction:

robots.txt file is used to restrict search engine crawlers from indexing selected areas of the website.

4.1.11.2 Dependency:

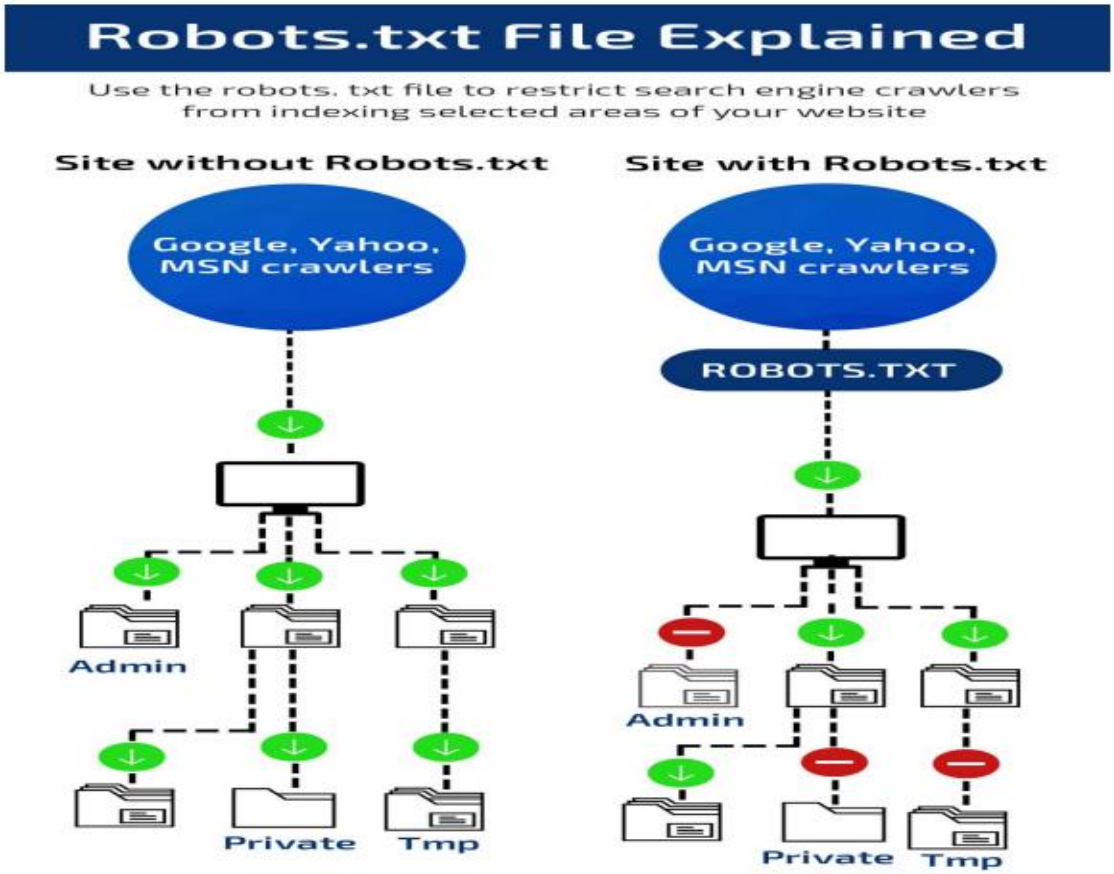


Figure 24: robots.txt architecture

4.1.11.3 Process:

The robots.txt file guides the search engine robots for the web pages that a website owner does not want to be crawled. For instance, if a trader does not wish specific videos on his website to be listed by Google or other search engines, then they can be blocked easily with the use of a robots.txt file. You can go to your website and check if you have a robots.txt file by adding /robots.txt immediately after your domain name in the address bar at the top. The URL you enter should look like this:

http://www.mywebsite.com/robots.txt (with your domain name instead!) When you create this file for your website, the search engine robots will look at the robots.txt to crawl and index specific pages of your site. Here I made url with made made a request if response status is 200 then it's present.

4.1.11.4 Experiment Result:

URL	https://www.amazon.in/
Robots Text Found	
Link	https://www.amazon.in/robots.txt

Figure 25: robots.txt Analysis

4.1.12 Keyword Analysis

4.1.12.1 Introduction:

Content matters, not only from an end user perspective, but also from a search engine perspective. The words used on a webpage, including what type (keywords or stop words), how they are used (alone or within phrases) and where they are used (link text or non-link body text), can all influence the value of the page in search.

4.1.12.2 Process:

The **Keyword Density Analysis** counts text found in the on-page body text, as well as text used within and heading tags. As a result, you can use this free keyword density tool to see which words and phrases are already used frequently on the page and mirror those words and phrases in the most valuable metadata tags, such as, heading tags, and alt text to improve the keyword relevance.

After running the keyword density tool, you may also find that you need to tweak the content of the page to ensure the most important keywords are used with reasonable frequency–but not to the point of keyword stuffing, which can

generate search engine penalties. Use this free keyword density tool to gain insight to the page content you already have. The keyword density tool can also help guide your SEO site review efforts to getting what you actually want and need on the page.

Stop words are words that carry no keyword relevance value, and as a result, the search engines generally ignore them when crawling and indexing a page.

4.1.12.3 Dependency:

After using the keyword density analysis tool, you may wish to make some changes accordingly to your website's content. While stop words confer no keyword relevance value, be sure your on-page content is written in normal language intended for human readers. Overly-optimized text that is little more than a keyword dump list will be recognized as such by search engine crawler algorithms as human reader unfriendly and can adversely affect a page's ranking. While it is good to take pointers based on the results of the free keyword density tool, always write your text content in natural, well-formed language specifically for your targeted human audience. Search engine algorithms value this type of content the most.

4.1.12.4 Experimental Results:

Figure 26: Keyword Density Analysis Algo 1

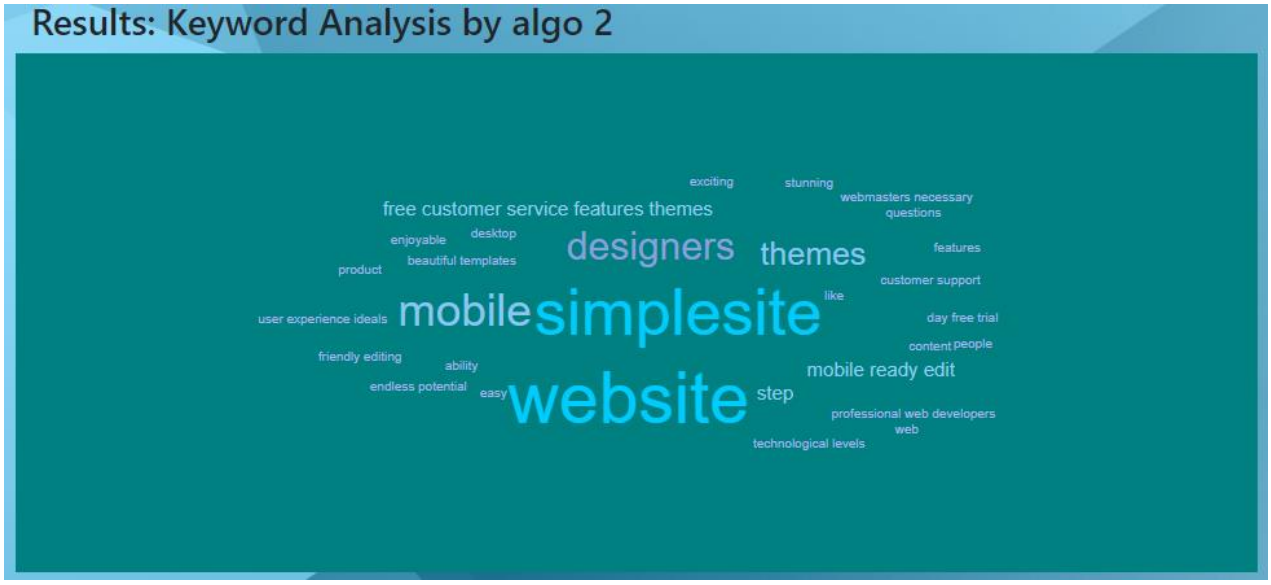


Figure 27: Keyword Density Analysis Algo 2

4.1.13 Competitive Keyword Analysis

Here we have taken the top five google search results based on search Keyword provided and the target site as a input and comparing the Keywords based and visually represented the word cloud. So that one can understands the keyword density of top five vs target site and based on that result we can update the keyword of target website to get better ranking.

4.1.13.1 Experimental Results:

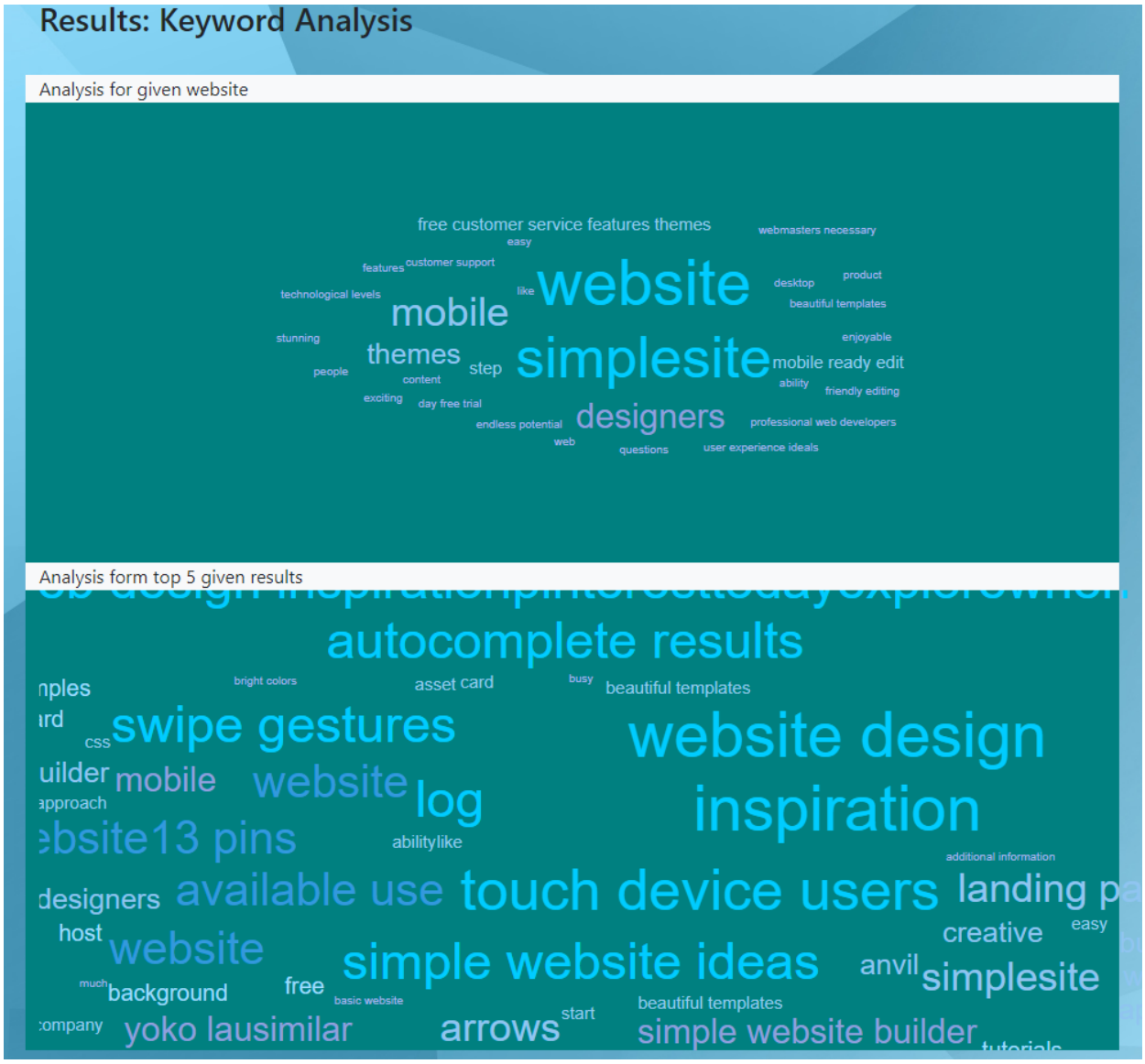


Figure 28: Competitive Keyword Analysis

4.1.14 Competitive Metatag Analysis

Here we have taken the top five google search results based on search Keyword provided and the target site as a input and comparing the Meta tag based and visually represented the table. So that one can understands the Meta tag details of top five vs target site and based on that result we can update the Metatag of target website to get better ranking.

4.1.14.1 Experimental Results:

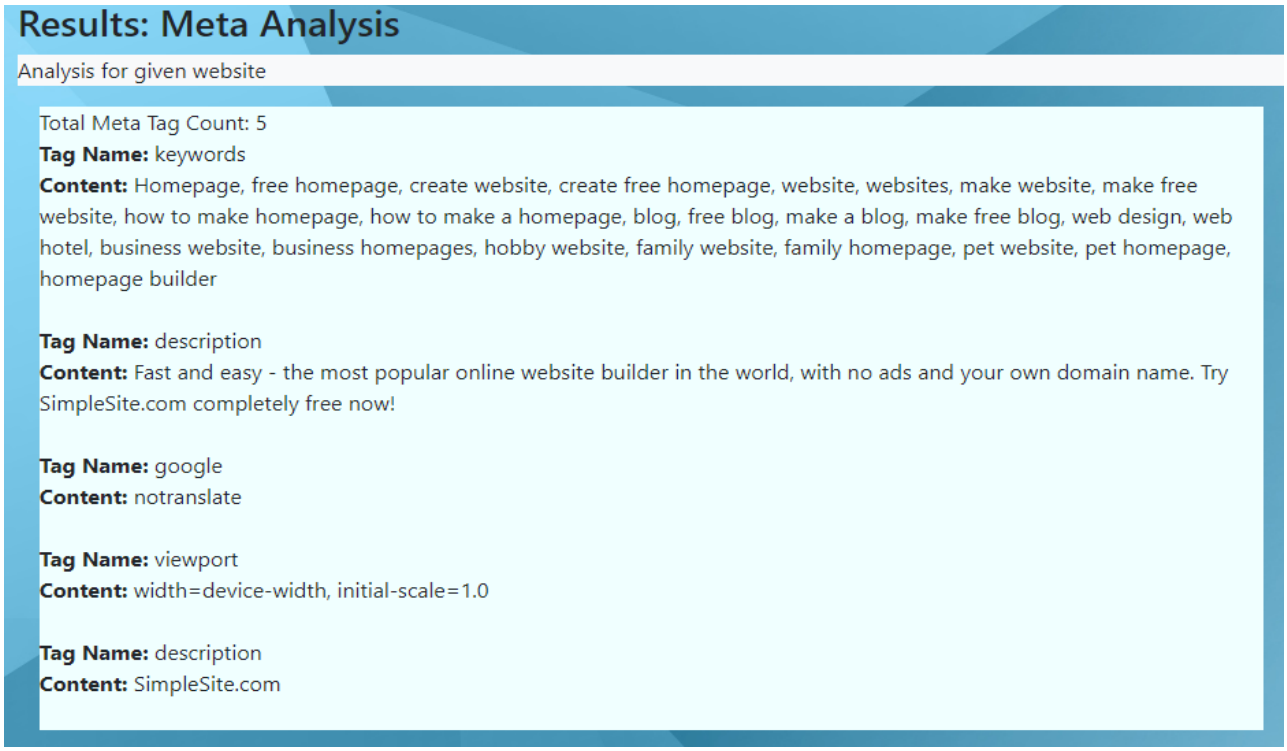
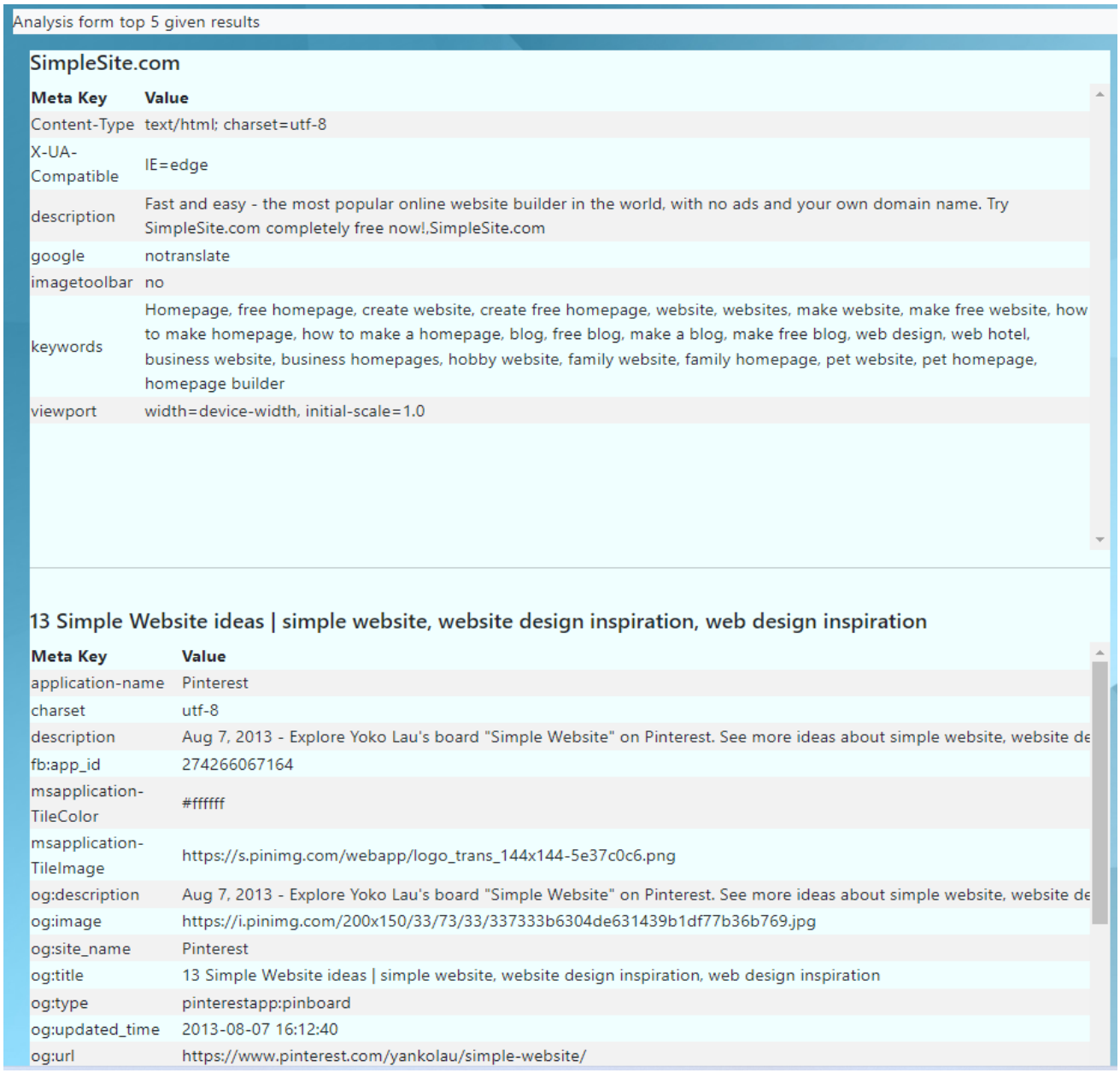


Figure 29: Competitive Meta tag Analysis (1)



Analysis form top 5 given results

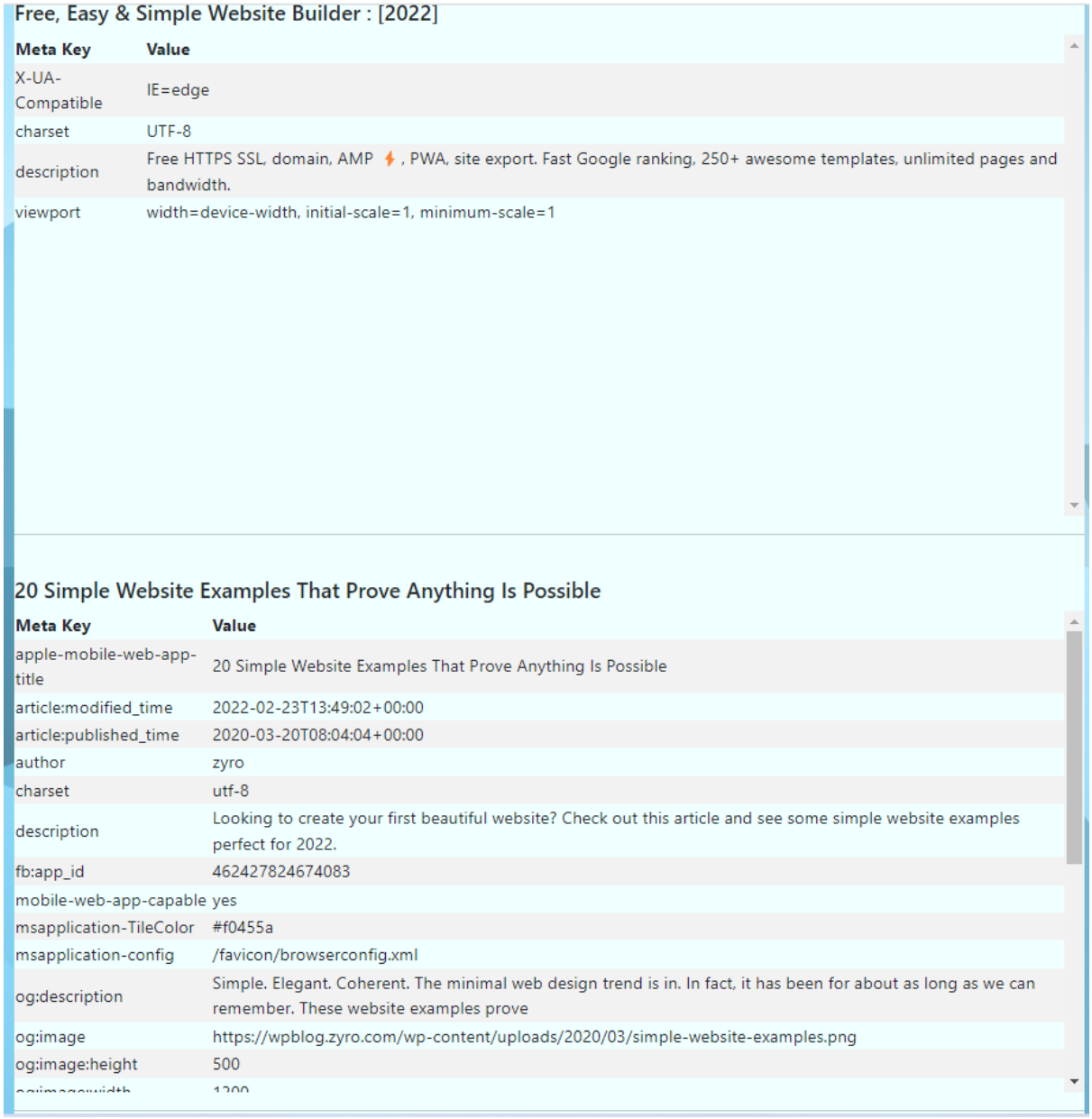
SimpleSite.com

Meta Key	Value
Content-Type	text/html; charset=utf-8
X-UA-Compatible	IE=edge
description	Fast and easy - the most popular online website builder in the world, with no ads and your own domain name. Try SimpleSite.com completely free now!,SimpleSite.com
google	notranslate
imagetoolbar	no
keywords	Homepage, free homepage, create website, create free homepage, website, websites, make website, make free website, how to make homepage, how to make a homepage, blog, free blog, make a blog, make free blog, web design, web hotel, business website, business homepages, hobby website, family website, family homepage, pet website, pet homepage, homepage builder
viewport	width=device-width, initial-scale=1.0

13 Simple Website ideas | simple website, website design inspiration, web design inspiration

Meta Key	Value
application-name	Pinterest
charset	utf-8
description	Aug 7, 2013 - Explore Yoko Lau's board "Simple Website" on Pinterest. See more ideas about simple website, website de
fb:app_id	274266067164
msapplication-TileColor	#ffffff
msapplication-TileImage	https://s.pinimg.com/webapp/logo_trans_144x144-5e37c0c6.png
og:description	Aug 7, 2013 - Explore Yoko Lau's board "Simple Website" on Pinterest. See more ideas about simple website, website de
og:image	https://i.pinimg.com/200x150/33/73/33/337333b6304de631439b1df77b36b769.jpg
og:site_name	Pinterest
og:title	13 Simple Website ideas | simple website, website design inspiration, web design inspiration
og:type	pinterestapp:pinboard
og:updated_time	2013-08-07 16:12:40
og:url	https://www.pinterest.com/yankolau/simple-website/

Figure 30: Competitive Meta tag Analysis (2)



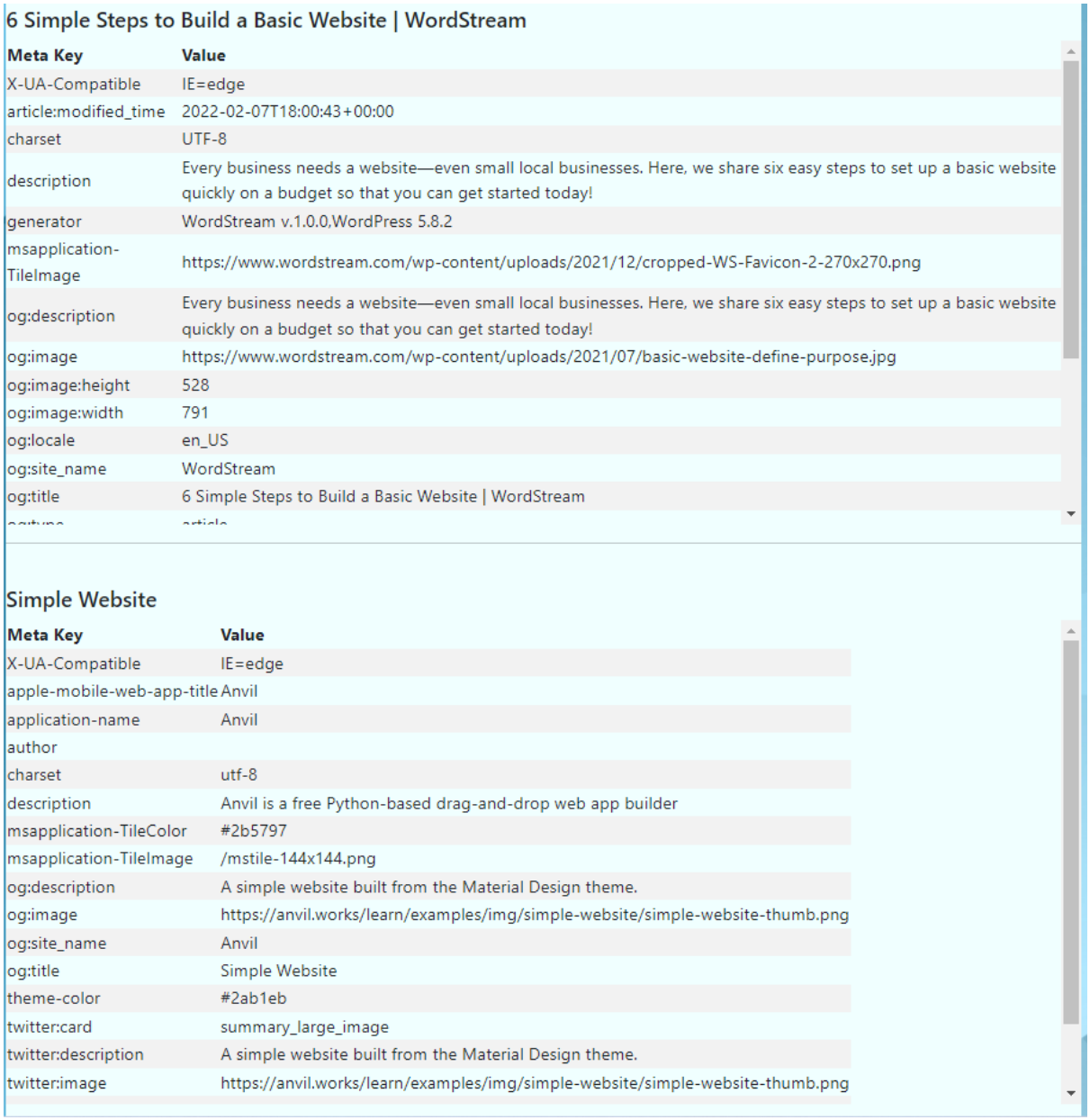
Free, Easy & Simple Website Builder : [2022]

Meta Key	Value
X-UA-Compatible	IE=edge
charset	UTF-8
description	Free HTTPS SSL, domain, AMP ⚡, PWA, site export. Fast Google ranking, 250+ awesome templates, unlimited pages and bandwidth.
viewport	width=device-width, initial-scale=1, minimum-scale=1

20 Simple Website Examples That Prove Anything Is Possible

Meta Key	Value
apple-mobile-web-app-title	20 Simple Website Examples That Prove Anything Is Possible
article:modified_time	2022-02-23T13:49:02+00:00
article:published_time	2020-03-20T08:04:04+00:00
author	zyro
charset	utf-8
description	Looking to create your first beautiful website? Check out this article and see some simple website examples perfect for 2022.
fb:app_id	462427824674083
mobile-web-app-capable	yes
msapplication-TileColor	#f0455a
msapplication-config	/favicon/browserconfig.xml
og:description	Simple. Elegant. Coherent. The minimal web design trend is in. In fact, it has been for about as long as we can remember. These website examples prove
og:image	https://wpblog.zyro.com/wp-content/uploads/2020/03/simple-website-examples.png
og:image:height	500
[illegible]	1200

Figure 31: Competitive Meta tag Analysis (3)



6 Simple Steps to Build a Basic Website | WordStream

Meta Key	Value
X-UA-Compatible	IE=edge
article:modified_time	2022-02-07T18:00:43+00:00
charset	UTF-8
description	Every business needs a website—even small local businesses. Here, we share six easy steps to set up a basic website quickly on a budget so that you can get started today!
generator	WordStream v.1.0.0,WordPress 5.8.2
msapplication-TileImage	https://www.wordstream.com/wp-content/uploads/2021/12/cropped-WS-Favicon-2-270x270.png
og:description	Every business needs a website—even small local businesses. Here, we share six easy steps to set up a basic website quickly on a budget so that you can get started today!
og:image	https://www.wordstream.com/wp-content/uploads/2021/07/basic-website-define-purpose.jpg
og:image:height	528
og:image:width	791
og:locale	en_US
og:site_name	WordStream
og:title	6 Simple Steps to Build a Basic Website | WordStream

Simple Website

Meta Key	Value
X-UA-Compatible	IE=edge
apple-mobile-web-app-title	Anvil
application-name	Anvil
author	
charset	utf-8
description	Anvil is a free Python-based drag-and-drop web app builder
msapplication-TileColor	#2b5797
msapplication-TileImage	/mstile-144x144.png
og:description	A simple website built from the Material Design theme.
og:image	https://anvil.works/learn/examples/img/simple-website/simple-website-thumb.png
og:site_name	Anvil
og:title	Simple Website
theme-color	#2ab1eb
twitter:card	summary_large_image
twitter:description	A simple website built from the Material Design theme.
twitter:image	https://anvil.works/learn/examples/img/simple-website/simple-website-thumb.png

Figure 32: Competitive Meta tag Analysis (4)

4.1.15 Website Ranking on Google

4.1.15.1 Introduction:

When Google was launched back in 1998, they introduced a mechanism for ranking web pages that was radically different from how the established search engines at the time worked.

Up to then, most search engines relied exclusively on content and meta data to determine if a webpage was relevant for a given search. Such an approach was easily manipulated, and it resulted in pretty poor search results where the top ranked pages tended to have a lot of keywords stuffed in to the content.

Google radically shook things up by introducing PageRank as a key ranking factor.

Content still mattered to Google, of course, but rather than just look at which webpage had the keyword included most often, Google looked at how webpages linked to one another to determine which page should rank first.

4.1.15.2 Process:

Here we are taking site URL as input and in backend we are comparing the top 1000 site domain based on keyword. If there is a match in domain, we calculated the Page rank and Site rank based on that.

4.1.15.3 Experimental Results:

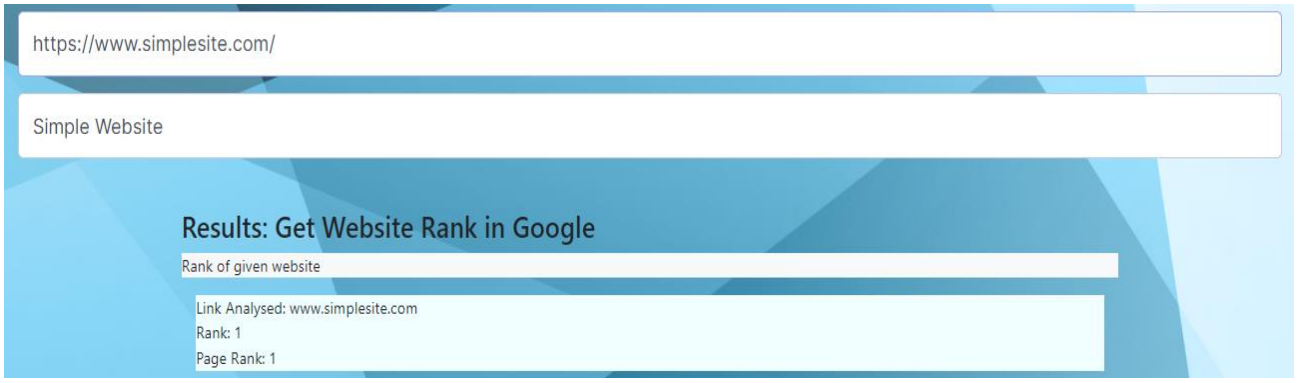


Figure 33: Website Ranking Results

4.1.16 Website Crawling

4.1.16.1 Introduction:

A web crawler, spider, or search engine bot downloads and indexes content from all over the Internet. The goal of such a bot is to learn what (almost) every webpage on the web is about, so that the information can be retrieved when

it's needed. They're called "web crawlers" because crawling is the technical term for automatically accessing a website and obtaining data via a software program.

These bots are almost always operated by search engines. By applying a search algorithm to the data collected by web crawlers, search engines can provide relevant links in response to user search queries, generating the list of webpages that show up after a user types a search into Google or Bing (or another search engine) [23].

A web crawler bot is like someone who goes through all the books in a disorganized library and puts together a card catalog so that anyone who visits the library can quickly and easily find the information they need. To help categorize and sort the library's books by topic, the organizer will read the title, summary, and some of the internal text of each book to figure out what it's about.

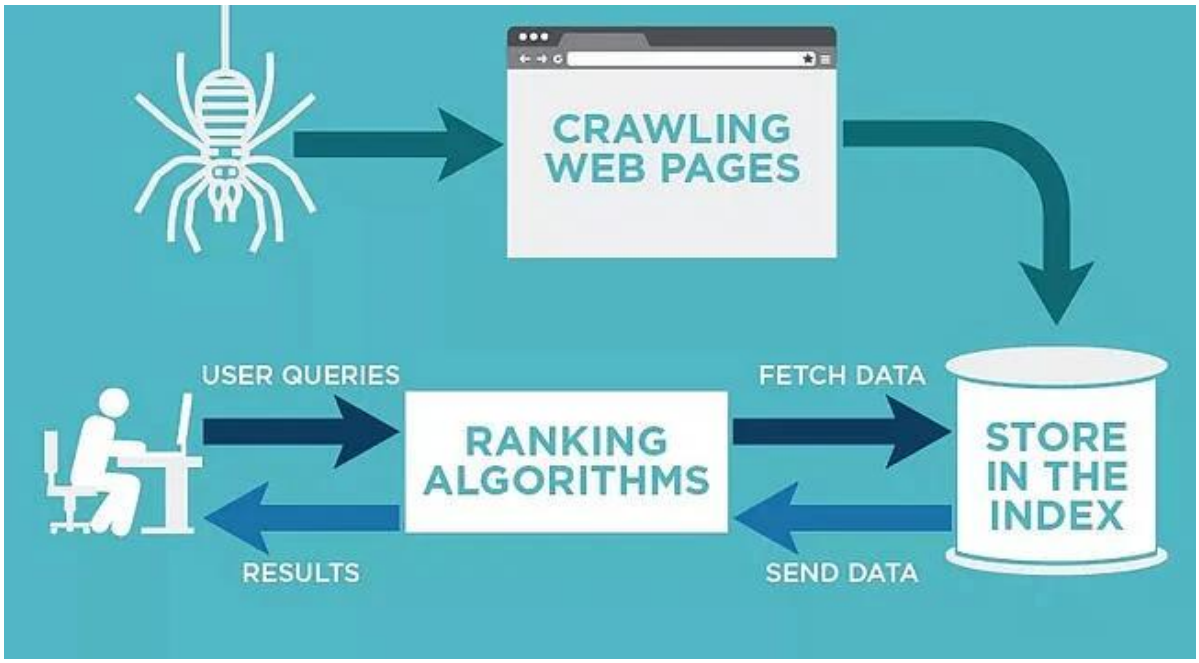


Figure 34: Website crawl Architecture

4.1.16.2 Dependency:

The Internet is constantly changing and expanding. Because it is not possible to know how many total webpages there are on the Internet, web crawler bots start from a seed, or a list of known URLs. They crawl the webpages at those URLs first. As they crawl those webpages, they will find hyperlinks to other

URLs, and they add those to the list of pages to crawl next.

Given the vast number of webpages on the Internet that could be indexed for search, this process could go on almost indefinitely. However, a web crawler will follow certain policies that make it more selective about which pages to crawl, in what order to crawl them, and how often they should crawl them again to check for content updates.

It is the discipline of readying content for search indexing so that a website shows up higher in search engine results.

If spider bots don't crawl a website, then it can't be indexed, and it won't show up in search results. For this reason, if a website owner wants to get organic traffic from search results, it is very important that they don't block web crawler bots.

4.1.16.3 Process:

Here we are taking a site and number of pages to be crawl as an Input, based on the input system is crawling and fetch different link and show us the results.

4.1.16.4 Experimental Results:

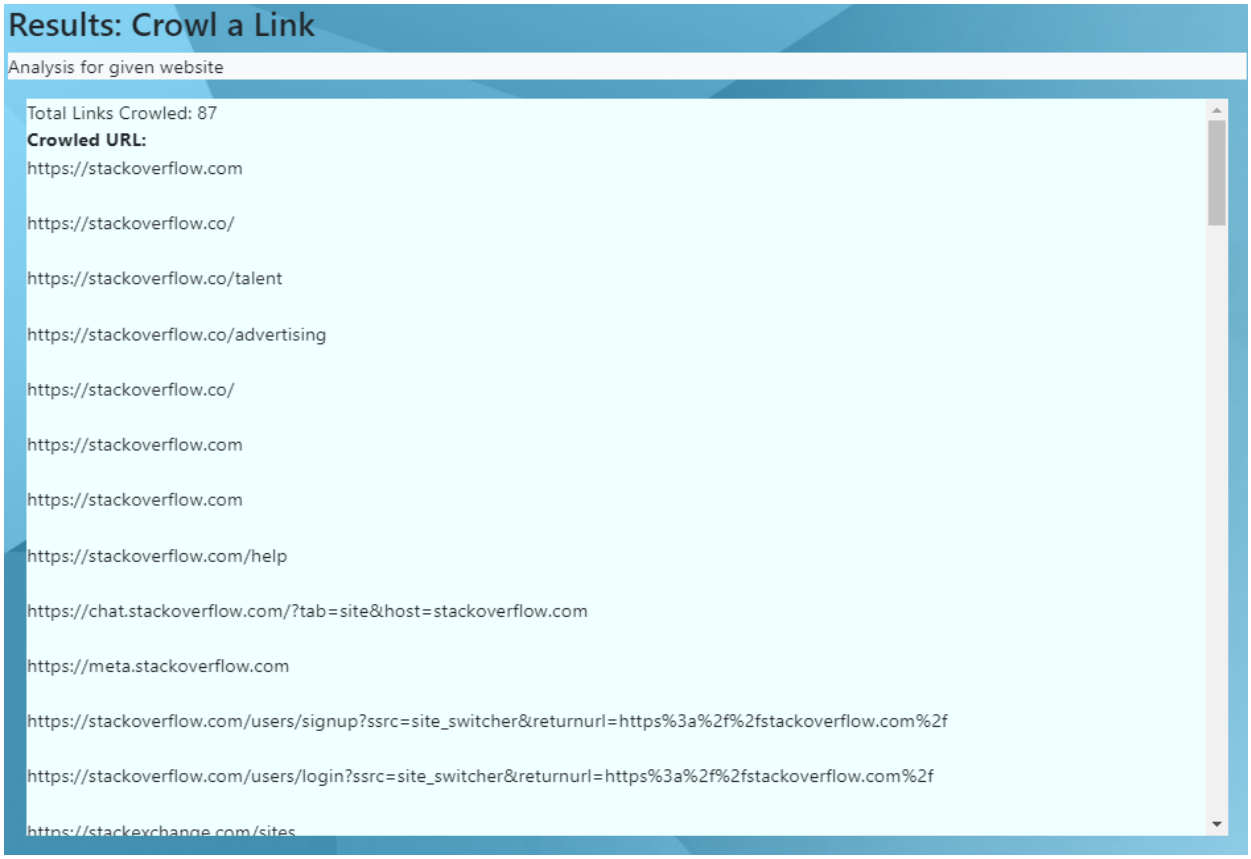


Figure 35: Website crawl Analysis

4.1.17 SEO Audit

An SEO audit is the process of analyzing how well your web presence relates to best practices – it is the first step to creating an implementation plan that will have measurable results.

The purpose of the audit is to identify as many foundational issues affecting organic search performance as possible. The SEO audit will reveal:

- Technical SEO issues
- Website structure issues
- On-page SEO issues
- Potential off-site problems
- User experience issues

- Content gaps and opportunities
- Competitive marketplace insights

An audit is a standard procedure that should occur on a regular basis – It is essentially a “health check” for your website.

4.1.17.1 Dependency:

First and foremost, an audit should be comprehensive. It should cover both structural and content components affecting your SEO visibility. It should provide a “big picture” view of what is happening in your current state. Any missing pieces could result in unnecessary or improper recommendations.

Your audit should be easy to understand. As a business owner, you should be able to connect the dots as to how SEO issues are affecting your online priorities, goals, or revenue. Any and all recommendations should clearly ladder up to your overarching business objectives.

Finally, your SEO audit recommendations should be actionable. There should be a clear path to completion; prioritized with projected impact and effort associated with each recommendation. The output of any SEO audit should accurately convey an easy-to-follow roadmap.

4.1.17.2 What Is Analysed in an SEO Audit

While SEO audits may vary slightly, each one should analyze basic technical SEO “table stakes” such as sitemaps, server errors, and metadata. At a high level, an audit should cover off on accessibility, indexation, and optimization. In an SEO audit at Bold Orange, our experts analyze over 193 different SEO elements across the following categories:

Technical Audit

- Indexation analysis
- Status Codes
- Redirects
- Page Speed
- URL Structure
- Robots.txt

- XML Sitemap
- Canonical Tags
- Duplication
- Crawlability
- Legacy Domain Issues
- Off-page Analysis
- Mobile SEO analysis
- International SEO analysis

On-Page Audit

- Site Content Structure
- Keyword Research
- Page Copy Theme Analysis
- Keyword Use
- Meta Data Analysis (Page Titles, Meta Descriptions, Heading Tags)
- Schema
- User Experience
- Images & Video
- Internal Linking & HTML Sitemap

Server Log Audit

- Crawl Budget Analysis
- User Agent Analysis

The information captured during the auditing process allows our team to make accurate, impactful recommendations for your specific website and situation.

4.1.17.3 Experimental Results:

Figure 36: Website Audit and SEO Analysis

Chapter 5

Tools and Technology Used in SEO Project

1. **VS Code**
2. **Anaconda**
3. **Spyder**
4. **Angular 13**
5. **Python 3.7**
6. **HTML, CSS**
7. **Typescript**
8. **GIT**

Chapter 6

Conclusion

Search engine optimization is a hard and tedious job, it can take a large amount of your time, there are too many pieces of the puzzle that you need to piece together. I once read "Promoting a site that writers on the web are unlikely to link to is as deadly as creating a fantastic website no one will see." I think the writer said it all in this simple sentence. There is no use in making the best website in your niche when you are not promoting it.

The search engines are getting more and more advanced in determining how and what webmasters are doing to get the rankings.

SEO is the act of modifying a website to increase its ranking in organic (v/s paid), crawler-based listings of search engines.

6.1 Future Scope:

Google's all algorithms are focused on improving the user experience. If you love your users, Google will love your website. If you have ever cared to closely look at the Google's algorithms, you already know that they will rank a website that takes care of the user's needs first.

Initially, SEO was considered to be a profile for IT department. As it was believed that SEO involves too much of technicalities. But if you understand SEO or have worked as an SEO professional, you would know it is an art to its core. The SEO professionals today understand that mere technical knowledge is not enough in this user-friendly virtual world. A creativity is required for an SEO professional to catch the interest of a user [24].

Ok! So, you know this now that now mere technical knowledge is not enough. You need to be aware of what the user is looking for. It is obvious that when a user makes a search, he is looking for something. If your site is substantial enough to provide him the solution. Then Google will recognize your site to be an authoritative source of information.

So as mentioned previously, in the past few years we saw keywords to be the

sole of SEO. However, with each passing year, Google made it very clear that the primary focus remains to be on the user experience. The annual graph of the changes made in Google's algorithm lay an extra emphasis on serving to what the user is looking for. The SEO professionals back then made attempt to find keywords of little relevance and stuff them into the blog.

The present SEO scenarios also suggest that the readability will have a lot to do with the engagement factor [25]. If your content is not able to hold the interest of the users, it is much likely to be pushed down in the search engine result pages.

Few of the Important factor in SEO future is:

- AMP (Accelerated Mobile Pages) will be an important ranking factor.
- Artificial intelligence will take care of the SERPs.
- Regularly update content based on website and trending keywords.
- Rankbrain is already in action. It is the machine system of filtering out websites on the basis of Google's metrics. Though it has many subfactors attached to it, it is one of the top three metrics of SEO rankings.

References:

[1] Beel, Jöran and Gipp, Bela and Wilde, Erik. "Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar and Co.". Journal of Scholarly Publishing. 2010; pp. 176–190.

[2] Bhandari RS, Bansal A. Impact of Search Engine Optimization as a Marketing Tool. Jindal Journal of Business Research. 2018;7(1):23-36. DOI: https://doi.org/10.1177/2278682117754016

[3] Google's 200 Ranking Factors: The Complete List (2022) by Brian Dean. Available online: https://backlinko.com/google-ranking-factors (accessed on 10 January 2022)

[4] Shraddha Londhe, Dr. Hemant Deshmukh, Dr. Rajendra Gode ," India Review Paper on Search-Engine Optimization". 2017.

[5] Stefanie Olsen. "Judge dismisses suit against Google". CNET. Retrieved May 10, 2007.

[6] Ziakis C, Vlachopoulou M, Kyrkoudis T, Karagkiozidou M. Important Factors for Improving Google Search Rank. Future Internet. 2019; 11(2):32. https://doi.org/10.3390/fi11020032

[7] Gupta, S.; Rakesh, N.; Thakral, A.; Chaudhary, D.K. Search engine optimization: Success factors. In Proceedings of the 2016, Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), India, 22–24 December 2016; pp. 17–21.

[8] Bing Webmaster Guidelines. Available online: https://www.bing.com/webmaster/help/webmaster-guidelines-30fba23a(accessed on 10 October 2021)

[9] Brin, S. and Page, The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems. 1998; 30, 1–7 , 107–117

[10] Veglis A, Giomelakis D. Search Engine Optimization. Future Internet. 2020; 12(1):6. https://doi.org/10.3390/fi12010006

[11] Zhang, J. and Dimitroff, The impact of metadata implementation on webpage visibility in search engine results (Part II). Information Processing & Management. 41, 3 (May 2005), 691–715. DOI: https://doi.org/10.1016/j.ipm.2003.12.002.

[12] Zhang, J. and Dimitroff, The impact of webpage content characteristics on webpage visibility in search engine results (Part I). Information Processing & Management. 41, 3 (May 2005), 665–690. DOI: https://doi.org/10.1016/j.ipm.2003.12.001

[13] Evans, M.P. Analysing Google rankings through search engine optimization data. Internet research. 17, 1 (2007), 21–37.

[14] Malaga, The Value of Search Engine Optimization: An Action Research Project at a New E-Commerce Site. Journal of Electronic Commerce. R 2007;

[15] Wang, F. et al. An empirical study on the search engine optimization technique and its outcomes.2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)(Aug. 2011), 2767–277

[16] Gasparotto, M. Search Engine Optimization for the Research Librarian: A Case Study Using the Bibliography of U.S. Latina Lesbian History and Culture. Practical Academic Librarianship: The International Journal of the SLA Academic Division. 4, 1 (Jun. 2014), 15–34.

[17] Karyotakis M-A, Lamprou E, Kiourexidou M, Antonopoulos N. SEO Practices: A Study about the Way News Websites Allow the Users to Comment on Their News Articles. Future Internet. 2019; 11(9):188. https://doi.org/10.3390/fi11090188

[18] Lee, S. et al. Search engine optimization: A case study using the bibliographies of LG Science Land in Korea. Library Hi Tech. 34, 2 (2016), 197–206

[19] Bala, Madhu and Verma, Deepak, A Critical Review of Digital Marketing. International Journal of Management, IT & Engineering, 2018; 8(10), 321–339., Available at SSRN: https://ssrn.com/abstract=3545505

[20] Zhang, S. and Cabage, N. Search Engine Optimization: Comparison of Link Building and Social Sharing. Journal of Computer Information Systems. 57, 2 (Apr. 2017), 148–159. DOI: https://doi.org/10.1080/08874417.2016.118344

[21] Rutz, O.J.; Bucklin, R.E. Paid search advertising. In Advanced Database Marketing; Routledge: London, UK, 2016; pp. 251–268.

[22] Su, A.-J.; Hu, Y.C.; Kuzmanovic, A.; Koh, C.-K. How to improve your search engine ranking: Myths and reality. ACM Trans. Web 2014,8, 1–25

[23] Al-Mukhtar, Firas & Hamad, Nawzad & Kareem, Shahab. SEARCH ENGINE OPTIMIZATION: A REVIEW. Journal of Applied Computer Science Methods. 2021. 17. 69-79. 10.23743/acs-2021-07.

[24] Mahmood, Nawzad. SEARCH ENGINE OPTIMIZATION. 2021;
DOI: https://doi.org/10.23743/acs-2021-07

[25] Fisenko, Tetiana & Baliun, Oksana & Grigorenko, Anna. SEO AS A METHOD OF WEBSITE PROMOTION. Věda a perspektivy. 2022; 10.52058/2695-1584-2022-2(9)-260-271.