Large-Scale Text Search Engine

M.Niwaes
Dept of CSE with AIML, SRM Institute of Science and Technology Tiruchirappalli,
niwaesmathavan@gmail.com

V.J.Shree Shanth
Dept of CSE with AIML, SRM Institute of Science and Technology Tiruchirappalli,
shreeshanth@gmail.com

P.Jaya Ruban
Dept of CSE with AIML, SRM Institute of Science and Technology Tiruchirappalli, rubanjaya@gmail.com

S.Sakthivel
Dept of CSE, SRM Institute of Science and Technology Tiruchirappalli,
Sakthivel.solaimuthu84@gmail.com

Abstract---The Search Engine has a critical role in presenting the correct pages to the user because of the availability of a huge number of websites, Search Engines such as Google use the Page Ranking Algorithm to rate web pages according to the nature of their content and their existence on the world wide web. SEO can be characterized as methodology used to elevate site keeping in mind the end goal to have a high rank i.e., top outcome. In this paper the authors present the most search engine optimization like (Google, Bing, MSN, Yahoo, etc.), and compare by the performance of the search engine optimization. The authors also present the benefits, limitation, challenges, and the search engine optimization application in business.

KEYWORDS: SEO, Page Ranking Algorithm, Google, Bing.

1. INTRODUCTION

Search engine optimization (SEO) is the mechanism by which a website or web page is improved to maximize the frequency and quantity of organic traffic from search engines (Kareem, 2009). Effective SEO means a web page is more likely to appear higher on the results page of a search engine (SERP). Google is the most popular search engine, but other search engines (Bing, Yahoo, DuckDuckGo, etc.) also have their own special web page crawling algorithms which return the top results of the search (Schwartz, 1998).

SEO is the process of helping to raise the rank of your website on Google and other search engines, thereby having your website in front of more users, growing company and making you a pioneer in the industry (Kareem & Okur, 2020a). Before you dive deep, it is incredibly critical that you have a good strategic strategy. Ranking for keywords is fine, but it is also critical, maybe even more so, to ensure that you meet your customers at each point of the purchasing process (Xu, Chen & Whinston, 2012). It's incredibly important to realize that SEO is not something you should take a half-assed approach to; before you dig in, you need to have a good strategy.

Ranking for keywords is fine, but it is also important to ensure that you meet your customers at each point of the purchasing process (Kareem & Okur, 2018). SEO tracking software such as Google Search Console make it easy to get quick insights on the results of page-specific or site-wide search engines (Kareem & Okur, 2020b). For a specific date set, 1 users can find statistics about which queries produce the largest volume of traffic, as well as the location a page rates for particular keywords. Although a page's search ranking might be the best predictor of SEO efficiency, when calculating SEO, there are several other success metrics that are valuable. Another website monitoring platform, Google Analytics, gives background for other indicators that could directly or indirectly influence the SERP rating of a blog. Such metrics include: page length, pages per visit, mobile traffic, bounce rate, visits returned (Schwartz, 1998).

2. LITERATURE REVIEW

There is still the very old discussion about who is best and why while addressing search engines. The internet is available through a multitude of engines, some arguably weaker than others, from Google and Bing, to Yahoo and DuckDuckGo. There are two main facets of the nature and responsibility of the search engine: its aesthetic appeal (how the results are for matted and displayed) and its consequences. By concentrating on the characteristics of the performance, as architecture is largely a secondary usability function (although still important) and contrasts the two main Google and Bing search engines Larry Page, Sergey Brin, launched Google in 1998 when Bing was developed as a succes sor to Microsoft's MSN Search, Windows Live Search and later Live Search in 2009. Google actually owns a 73.02 percent share of all users of the desktop search engine, while Bing only holds a 9.26 percent share (Schwartz, 1998).

Michael Basilyan, a senior program manager at Bing, reported in 2014 that "content quality" is one of the key priority areas of the page ranking phase of the search engine. In its rating estimation, Bing takes three attributes into consideration: the contextual importance of a website, the meaning, and the consistency of the content. Topical significance questions if a website page is linked to the query, i.e. "Does it address the query?" content analyzes contextual and historical users and asks specifics, such as whether the query might be about a common recent issue, what the user's physical location is, the search history of the user, etc. finally, and most notably, the standard of content explores three crucial questions: "Can we trust this content?" "Is the content helpful and sufficiently comprehensive?" and "Is the material easy to find and well presented?". While these are not the only variables taken into consideration by Bing in its page ranking system, they are what Bing finds to be the most important. In the other hand, for Google, it's impossible to rely on only a few key rating considerations. Google SEO consultants could name the most significant factors in the ranking algorithm with over 200+ ranking factors, such as keyword use, site structure, site speed, time spent on site, number of inbound links, and inbound connection quality. Google and Bing use many of the same page rating variables, quite clearly, but what separates the two rivals is the variations in how they use them. While the disparity between Bing and Google is not dramatic, certain characteristics are worth mentioning. While Google's search algorithm excels in matching synonyms and related terms with queries and data, Bing needs more precise matching of keywords to acquire precise search results. In addition, Bing continues to assign preference in page ranking to pages that are important to recent events. For eg, the first link on Google is the official Indian tourism site while searching for it in Google and Bing, while Bing shows news articles first; preceded the ranking of the official tourism website. However, if a website includes Flash content.

Due to difficulties linking to a single article, Google's algorithm appears to often rate pages with flash lower than their counterparts. Finally, unlike Google, Bing appears to take the domain age of a website further into account and has a bias towards. Bing and Google varied only marginally in the effects they showed while researching the algorithms and page outcomes of the two search engines. An analysis conducted by SurveyMonkey showed in a study by Search Engine Land that most users would choose search results branded as 'Google,' even though the results were directly created by Bing. Participants preferred Google searches 57 percent of the time and Bing 43 percent of the time, a figure slightly lower than the current market share for the two search engines, while no company name was imposed on the results. it comes down to a mixture of individual tastes and brand prejudice. Despite Bing's recent modifications in its search algorithms to the point that the page results of Bing and Google are impossible to discern, Google has become a household name that is well respected and consistent in its results and the importance of its page rankings for the majority of searches (Smith, 2010).

3.METHODOLOGY

Search Engine Optimization is a short version of the word SEO. The optimization process was usually designed to illustrate the search results carried out by offenders of search engines such as Google, Yahoo Overture, etc. The web pages of these websites are ranked in the top ranks. From the above, it can be inferred that, in short, SEO is a method of improving the search engine that will be able to produce the necessary search results available to users. Search Engine Optimization is known to be an efficient way to optimize the eminence and degree of user traffic by the intrinsic search possibilities for the given website or domain. This is regarded as a type of natural search. This can be divided into algorithmic and organic again. The higher the ranking given to it by the SEO, the greater the search for that particular website or domain. A frequent refresh of the contents and, in particular, axioms that will increase the traffic will help maintain the high-ranking rates. In collaboration with various practitioners and groups who try to collect understandable knowledge and the method to determine the importance of a given keyword optimal for the search inquiry, search engine optimization is a big industry. Crawling, indexing, sorting, measuring importance, and re trieving are some of SEO's essential tasks. From the point of view of the aforementioned discussion, it can be understood that the optimization of the search engine is a type of opera tion that will increase the eminence and degree of user traffic for a particular business intent of the organization's website. The search engine optimization function is based on the algorithm of the search engine. The search engine can have natural capability checking methods that can be of two types, namely algorithmic and organic scanning processes. Generally, a ranking is given to the website that, when it is checked for full time and is ranked among the first 10 queries, would formalize a high rate .

As it is noted that the first set of listings in the search engine is invariantly clicked by many of the customers, it is very important for company websites to face a hectic rivalry. The optimization of the search engine is known to be a mechanism designed to boost the visibility of the website in question.

Statistical Method	Automatic Method
Statistical methods require relevance judgments from experts and searchers to prioritize the web pages in the search engine's database	Automatic methods use the users' interaction with browsers to assess the quality of web documents.
Statistical methods require additional cost for expensive experts' judgments.	Automatic methods for search engine evaluation require low cost.
No real-time data is gathered.	Only real-time data are collected.
Long time period is required to evaluate the web pages.	No additional time period is required.
Limited to decisions of a small group of people	Web pages are evaluated and assigned numeric scores based on decisions from all the searchers.
Searchers' role becomes partial in relevance score computation.	Searchers play a vital role in deciding the usefulness of web documents.
Experts know their contribution in evaluation of web documents.	Searchers often do not know about the hidden judgments collection.

3.1 SELECTION OF SEARCH ENGINES

Yahoo, Google, Gigablast, AlltheWeb, Zworks, AltaVista, and Bing/MSN were the search engines chosen for comparison in this report. During the discovery process of the Web search engines to be analyzed, attention was paid to the inclusion of a number of search engines so that the results obtained could serve as a basis for assessing the search algorithm used by the different search engines. Some of the search engines chosen are not the most common or the most recognizable ones. Therefore, the findings of the analysis would educate users about their various capabilities and thereby theoretically improve the usage of search engines that perform better. In addition to Web records, certain search engines often index content stored on other Internet applications, such as chat groups and Gopher (a network that directs users to businesses that provide those goods and/or services), but this analysis considered only Web databases. Also, centralized web search engines such as CUSI (Configurable Unified Search Index) were not considered because they just compile and do not provide something new with current web information (Edosomwan & Edosomwan, 2021).

3.2 Test queries

On all search engines, ten search queries were planned for use. These queries were meant to assess different features that each search engine claims to provide, as well as to reflect varying degrees of difficulty of searching. For the sake of familiarity, the searches were often meant to fall into the information technology domain, so that the investigators could judge the search results as appropriate. In four classes, the ten queries were listed as follows:

a. Short queries:

What is data mining? (Query 1) Web Navigators (Query 2) Neural networks (Query 3) Evolution of processors (Query 4) Keyword searching (Query 5)

b. Boolean logic (AND/OR) queries:

Searching AND sorting (Query 6) Clustering OR clustering algorithm (Query 7)

c. Natural language queries:

Search the Internet using natural language (Query 8) How do I get the best search result on the Web? (Query 9)

d. Response time:

Answer time was determined by a stopwatch and was calculated as the interval between entering a search query and obtaining the first search results. To determine the response time, we picked one question from each category. The selected queries were: Query 1 (Group A), Query 6 (Group B), Query 8 (Group C) and Query 10 (Group C), respectively (Group D). It then determined the average response times for each search engine and for each question chosen (Edosomwan & Edosomwan, 2021).

e. Test environment

As the web browser for the analysis, Microsoft Internet Explorer was chosen because it is compatible with all the search engines selected and is locally the most commonly used browser. Two machines with different settings but with the same specifications were used: An Acer computer with an Intel Celeron M 440 CPU, 80 GB of hard drive (1.86 GHz) and 52 MB of DDR2 ram, and a Hewlett Packard computer (2.10 MHz) with an AMD Semipro SI-42 processor, 140 GB of hard drive and 1 GB of RAM. Those obtained from the Hewlett Packard machine are the findings shown. Results from the repeated exercise are not discussed because the results of the analysis have been similar and do not improve. Ideally, any question should be executed at the same time on all search engines, meaning that neither should have the benefit of being able to index the new page above the other one if a relevant page is added. That was not technically realistic for this analysis and so each question was checked on all the search engines on the same day within thirty minutes of each other. In order to be returned, all search engines that return an error of '404' (i.e., route not found) or '603' (i.e., server not responding) are noted. Return trips were made at varying periods of the day to prepare for the likelihood of daily maintenance downtime for the facility (Edosomwan & Edosomwan, 2021).

f. Precision

Accuracy was described for this analysis as the relevance of a search result to a search query and was calculated independently for the first ten search results by both investigators. In order to decide if it satisfied the intended outcome, we reviewed the quality of each obtained outcome, but did not attempt to read the full-text Web document by following the links given due to time considerations and variable connection reliability. A precision score was determined on the basis of the number of outcomes considered significant during the first ten collected (i.e., a score of 1 indicates that all ten search results were relevant and a score of 0.5 indicates that only five of the first ten results were relevant). We not only measured the average accuracy score for each question to determine the overall output of each search engine we tested, but also determined the average accuracy score for each search engine, based on all ten queries (Edosomwan & Edosomwan, 2021).

g. Working of search engine optimization

To grasp the idea in depth, the operation of the search engine is very important. The search engines typically optimize a specific web page whose very small and special main term or constraint is coined once after the optimization type is submitted, then ranked as the first few search results based on the sum of the search frequency. This will also increase the traffic as the customer performs a request for a specific word. The search engines carry out frequent updates to the material in order to help the pages at a comparable stage.

In two distinct ways, namely human power-driven, the search engines can be categorized and the other is a crawler-supported engine.Therefore, the study is carried out in a step-by-step phase in the row as identification of relevant key words, estimation of market strength, enhancement of the website, and creation of web page connections, acquiescence, inquiry, adjustment and reporting. Engine powered by Crawler: This sort of search engine is intended to send a spider to follow multiple sites and apply the spider's results to the search engine's indexes. In a return of the spider, the updating of the websites is done once as it gives the new details through looking. Therefore, when the quest in this engine begins, the entire index is scanned for matches with the searched word. Google is an example of the Crawler type of search engine. The index is a kind of bulky catalog that stores a copy of each web page uncovered by the spider and checks the index. Human driven search engines: a small representation of the article such as the name of the author or the name of the individual organization that will fit the engine directory will be needed for this sort of search engine. Through upgrading the new pages, this sort of search engine would not vary the ranking of a given domain, but instead incorporate the new stuff so that each website can be searchable. Look Smart is an example of this type of search engine. Hybrid Search Engine: Separate search engines are now using all the aforementioned approaches to improve web page content in the index referred to as 'hybrid search engines.' MSN and Yahoo are a case in point for this sort of search engine. It can be inferred from the above topic that, before beginning to customize a web, the operation of search engine optimization is very important. This would begin with a clear procedure flow, such as the identification of relevant key words, the determination of the level of competition, the enhancement of the page and the development of web page connections, approval, inquiry and alteration and documentation. Search engines can be assisted by crawlers, human power powered and hybrid search engines, among oth

er types. Each will confine the search process to a different method. A spider is sent to follow different places in the crawler system, and the resulting spider finds are added to the search engine indexes after a single term is provided in the search engine choices. In order for the quest to begin, the human guided method would require a short summary or the name of the author of the writings. The search engine hybrid form will show all the original search process systems so that the full search capability is given in this form (Mustafa, Yousif & Abdulqadir, 2019). Depending on the degree of relevance, the classification or rating of the websites may be given to the webpages of the crawler type of a search engine. The importance of matching the term with the key terms identified is accomplished by using an algorithm that differs with each search engine type, but the functioning of the algorithms is identical to: Tracing main constraints: The search engines scan the headlines, title tags or the opening of two snippets to trace the highest matching word with the text (Mohamed & Khoshaba, 2012). Main restriction frequency: The search engine tests the frequency of a given constraint in such a way that the term frequency predicts the accuracy of the web. Spam prevention: spam creation, such as click-through assessment and connection scrutiny etc., is not constrained by the optimization of the search engine (Hawezi, Azeez & Qadir, 2019).

h. SEO Benefits Search Engine

Optimization has acquired a lot of boom and the value has increased further with the increased Internet offenders (King, 2008). For a business, the optimization of search engines has several advantages: Both regionally and internationally, the axioms or primary restrictions outlined would encourage dominance in the viewer. For enterprises that run abroad, this would be very useful. When run with important and most appropriate axioms and key terms, the search engine would maximize the degree of traffic for the website of the specific company. The SEO aims to transform searchers' traffic to prospective customers and is therefore considered the best way to improve the business. After the optimization process, the visibility element of the company's website will begin. The customers are also aware of the exclusive services and goods of the business firm in question. The optimization of the search engine is found to be more functional and advantageous relative to any other form of traditional marketing. Compared to all other marketing types, the optimization of the search engine is successful in increasing immense earnings on investment returns. This would increase the company's revenue and earning (Kareem, 2009). The rating offered by the SEO would help to maintain the website of the firms for a very long time for comparison and is a very cheaper solution compared to other approaches (Abdalwahid, Yousif & Kareem, 2019). From the above discussion it can be understood that, the search engine optimization will provide lethal advantages to the businesses such as controlling the traffic volume, increased sales and revenues, high profits, more benefited way of advertising of the services and products of a company, cost effective, high range of visibility, global and local visibility, lesser capital for investment, etc. (Kareem, Yousif & Abdalwahid, 2020). From the aforementioned discussion, it can be understood that the optimization of the search engine would provide firms with lethal benefits, such as traffic flow management, improved sales and earnings, high profits, more beneficial advertisement of a company's services and goods, cost-effective, high exposure range, global and local visibility, reduced venture capital, etc.

i. Limitations of SEO

Search Engine Optimization has different challenges and limits (UKEssays, 2018). Few of these are discussed in the following manner: Limitation on idioms and key restrictions: the biggest limitation is met in the form of inadequacy of the same axioms and key constraints combined for the same domain with a single platform under search engine optimization (Amin, Shahab, Al Azzawi & Sivaram, 2018). This would be a big downside and will decrease the human traffic on that specific location. This is due to uncertainty over choosing a fitting axiom.

This is due to uncertainty about the collection of sufficient axioms. Competition limitation: Competition increases when two separate websites are protected by the same axiom or primary restrictions. In this case, in order to achieve a certain rank, websites are met with opposition. Subpage limitation: The subpages of the websites must still retain the axioms and restrictions that are in a position to be changed each time and remain updated in order to improve the ranking and increase the traffic (Berman & Katona, 2013). Lingually: Most conventional search engine optimizers run on a single language medium that can restrict the search to a specific geography.
Limitation of crawlability: After passing through millions and millions of webpages, the crawling nature will be constrained and thus the speed of the search will be impaired (Zhang & Cabage, 2016).Duplication limitation: In the search engine, the duplication of a single webpage happens when a specific page is submitted several times with similar content (Yang & Ghose, 2010).

j. SEO and Business

The advantages of new development sovertheconventional optimization of search engines are as follows:

The break through optimistically minimized the replication process of the original material of the web pages.
Increased pace of the crawler's search skill.
The essence of scalability has improved.
Enhanced effectiveness and traffic for the company.High scores are probable. It can be inferred from the above that the advantages of modern technologies on the conventional searchengineforbusinesspurposesaretoboosttraffic,decreaserep etition, speed ,scalability, enhance effectiveness, etc. (Smith,2010)

Fig 1:Diagram of SEO and BUsinees

3.RESULT

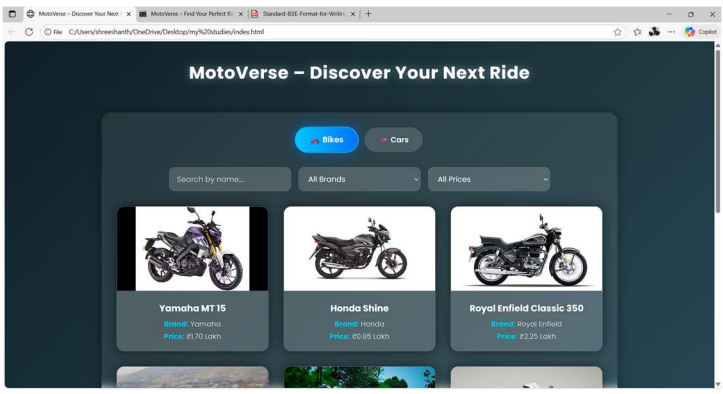

Fig 2a: results of the program page1

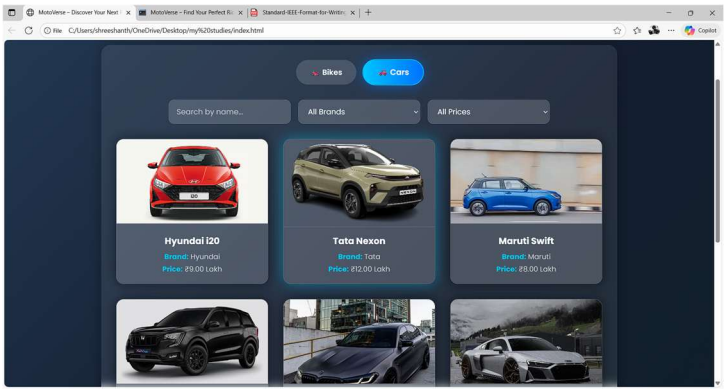

Fig 2b: results of the program page1

4.CONCLUSION

The benefit of search engine optimization is essentially the fact that it increases your website's popularity. Visibility is everything in modern business if you intend to go ahead. People ought to be able to locate you, because given the number of rivals, i.e., others who choose to be positioned for the same keywords, this is not a simple feat. It should be able to appreciate how this reflects your industry as you understand the value of exposure. You can see how other areas of industry are impacted, such as sales, reputations, etc., starting with the number of visitors to the website, which is the first one to change once you maximize exposure. If you ignore the benefits SEO provides, you will miss dramatically in terms of the three variables: advertisement, publicity and sales

.What you need to keep in mind is that improving search engines is a long-term task that will need months to demonstrate any signs of progress. No immediate strategies and resources are available that will help you achieve immediate results, regardless of what anyone says otherwise. The reports are more accurate and long-lasting, considering the time taken to spend in working on SEO. Some of the most commonly used software that can assist with multiple activities that are part of SEO, ranging from keyword study to website review, are the tools we suggest in this book. There are lots of other apps, both free and charged, and the latest ones that are being created at this time, so feel free to explore and discover the resources that you find easy to use and that can really help with the tasks as part of SEO that you are about to perform. Finally, as the internet has become such an expansive and competitive virtual environment, a comprehensive, tactical and precise endeavor is to get the best out of the potential a search engine offers.

REFERENCES

1.Abdalwahid, S. M. J., Yousif, R. Z., & Kareem, S. W. (2019). Enhancing approach using hybrid Pailler and RSA for information security in Big Data. Applied Computer Science, 15(4), 63–74. https: /doi.org/10.23743/acs-2019-30

2.Amin, S. M., Shahab, W. K., Al Azzawi, A. K., & Sivaram, M. (2018). Time series prediction using SRE-NAR and SRE- ADALINE. Journal of Advanced Research in Dynamical & Control Systems, 10(12), 1716–1726.

4.Berman, R., & Katona, Z. (2013). The role of search engine optimization in search marketing. Marketing Science, 32(4), 644–651.

5.Edosomwan, J., & Edosomwan, T. O. (2021). Comparative analysis of some search engines. South African Journal of Science, 106(11/12), 169. http: /doi.org/10.4102/sajs.v106i11/12.169

6.Hawezi, R. S., Azeez, M. Y., & Qadir, A. A. (2019). Spell checking algorithm for agglutinative languages "Central Kurdish as an example." In 2019 International Engineering Conference (IEC) (pp. 142–146). IEEE.

7.Kareem, S. W. (2009). Hybrid public key encryption algorithms for e-commerce [Master's thesis, Salahadin University].

8.Kareem, S., & Okur, M. C. (2018). Bayesian network structure learning using hybrid bee optimization and greedy search. Çukurova University, Adana, Turkey.

9.Kareem, S., & Okur, M. C. (2020a). Evaluation of Bayesian network structure learning using elephant swarm water search algorithm. In S. C. Shi (Ed.), Handbook of research on advancements of swarm intelligence algorithms for solving real-world problems (pp. 139–159). IGI Global.