IJARSCT

ISSN (Online) 2581-9429

IJARSCT

Impact Factor: 6.252

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 2, Issue 6, May 2022

Search Engine Algorithms

Anand Sharma[1], Vaibhav Chandekar[2], Vikas Damre[3], Shrutam Dhone[4], Kapil Patil[5]

Project Guide, Department of Information Technology[1]
Project Group Leader, Department of Information Technology[2]
Project Group Member, Department of Information Technology[3, 4, 5]
Shri Sant Gajanan Maharaj College of Engineering, Shegaon, Maharashtra, India

Abstract: *A search engine algorithm is a complicated method that search engines like Google, Yahoo, and Bing use to estimate the importance of a web page. There are about 150,000,000 active websites on the Internet, according to Netcraft, an Internet research firm. There would be no way to tell which of these sites are worth users' attention and which simply spam without search engines are. Search engines collect a lot of information, which allows them to quickly decide if a site is spam or contains useful information. Search engines reward relevant sites with high rankings, whereas spam and unrelated sites might obtain extremely low rankings. There are three components to a search engine algorithm: crawling, indexing, and ranking.*

Keywords: Algorithm, Search engine, Crawling, Indexing and Ranking

I. INTRODUCTION

On the web, there are numbers of thousands of articles ready to present data on a wide range of informative and fascinating themes. Search Engines are messengers that provide the same information to you should you need it. The sad fact is that while there are millions more pages waiting to mislead you with misleading info formed by the writer's taste and desire. Search Engines are indeed the saviours in this circumstance, as they prohibit these fraudulent internet sites from coming out to you. When a certain user types in a search engine the search query, all of the shown pages in the index that are judged relevant are identified, as well as the algorithm is also used for rank the important passages hierarchically as a result in a set. Each search engine's algorithm for ranking the most effective findings is different.

II. Literature Review

Web search engines (WSEs) are used by people for study, browsing, and amusement. Without the assistance of a WSE, conducting such actions manually would be impossible due to the vast number of Websites. Furthermore, the usefulness of WSEs is always improving. One can acquire many URLs containing the appropriate contents by just querying the WSE with a few keywords. WSEs, on the other hand, aren't just confined to returning a list of URLs.

The WSE does a search and processes and stores a query (unstructured data). The WSE will save a chronology, the URL chosen by the user, and any other potential information gathered about the consumer throughout the search along with the query. The query log belongs to all this added meta data, and also the query itself. The WSEs process and analyse streams of query logs in order to construct and improve user profiles. Users should expect a better service as a result of this.

R. Khalil and N.A.K. Muhammad suggested a revolutionary technique that improves the efficiency of user search data. This approach establishes a link between searches, documents, and the user query. Take into account the semantic document structure as well as the user inquiry. The findings of the proposed approach are superior than those of earlier approaches. A modified page ranking algorithm was proposed by R. Seema and G. Upasana. On the ground of incoming visit links on pages, the new algorithm calculates page rank. The VOL algorithm was introduced PR algorithm that outperforms the original.

The results reveal that VOL is superior to the original PR method, and that pages with more incoming link visits have a higher rank value than pages with fewer visits. A method for determining the link-visit counts of Web pages is also shown, as well as a comparison of VOL and the PR algorithm.

J. Ayush offers a new technique for determining web page rank based on several parameters. Modified HITS is a suggested algorithm that improves on the HITS algorithm. It was created by extending the capabilities of the HITS algorithm. Six parameters are taken into account and used to calculate the web page rank.

DOI: 10.48175/568

Quinn Norton, (ISM). In introduced a new method for indexing online pages utilising ISM, in which the significance of the search term is translated and subsequently web sites are indexed depending on the interpretation. Existing approaches and limitations with several algorithms in use for network analysis, such as PR, WPR, HITS, and CLEVER algorithm, were also examined.

Proposed Content Based Hidden Web Ranking Algorithm, "Xenon web crawling initiative: security impact assessment (PIA) summary" (CHWRA). There are four different attributes in the proposed method. This strategy seeks to account for all factors that influence the popularity of a website, whether directly or indirectly. This function returns an ordered result set from a Hidden web search. The CHWRA algorithm produced the expected outcome.

III. Comparative Study

Crawling:

This section will go over each of the previously listed categories, selecting an example from each and elaborating on the traits and architecture that distinguish them from other approaches. A thorough examination of the instruments will provide the information needed to make conclusions from the research.

Crawlers: Visual vs. Programmatic

On the web, there are several "visual web scraper/crawler" programmes that will able to crawl pages and arrange data into therowsas well ascolumns based on the user's needs. The programming ability level is necessary to set the crawler is one of the main key differences between a visual and a classic crawler. The most recent generation of "visual scrapers" eliminates the most of the programming knowledge required to construct and launch a crawl to scrape online data.

The user "teaches" a chunk of the technology of crawler, which then explores trends in semi-structured sources of data using the visual scraping/crawling method. Using a browser to highlight data and educate columns and rows is the most common way to teach a visual crawler. Whereas the technology isn't really fresh, entrepreneurs and finished are continuing to invest and expand in the area.

The another name of web crawler is as spider bot or spider as well and most commonly known as crawler, it is an bot that is the bot of the internet that browses World Wide Web (WWW) inside out and conventionally used by search engine to index content (web spidering).

Usually search engine uses online spidering or crawling softwares and a few other websites to refresh the web contents or indices of many other website’s web contents. Web crawlers always save pages to be prepared by the search engine, that basis the pages that’s why users can able to find the information quickly and accurately.

Crawlers use the resources on the visited a systems and frequently visit the sites without being asked. When big collections of pages are accessed, load, schedule’s issues, and "politeness" come to play. There are mechanisms in place for the public sites that do not want to get crawled to notify the crawling agent. A robots.txt file, for example, it is possible to direct bots to index only a certain part of the website or even not to index it at all

The amount of pages on the Internet is enormous, and even the most powerful crawlers can't create a complete index. As a result,the World Wide Web is in early year, before 2000, It was difficult to find relevant search results from the search engines. Relevant results are now practically instantaneously available.

Indexing

By using indexing techniques, by reducing the number of disc accesses required we can improve database performance to run the query. Basically, it is a method of storing and retrieving data from the database that can be used to quickly find and access data.Few of the database columns is used in order to create indexes.

There are two columns in the table: the Search key and the Candidate key. The Search key always contain a duplicate of either of the table's two keys.

Secondly, the Data Reference or Pointer columns. These columns provide a collection of pointers containing addresses of disc blocks that contain the key value(s) for the particular key, and these are the third and final columns.

The indexing has a number of characteristics:

There are several types of access available, such as valuedbased searches, range searches, and so on, that you can make use of.

The time which takes to locate a particular element of data or the data combination pieces is commonly referred to as access time.

Insertion Time: Basically, it represents the time that which take to locate the right space and put the new data into it.

Deletion Time: It is the amount of time which take to locate and delete the item while updating the index structure at the same time.

The term "space overhead" refers to the extra space that the index must occupy as a result of the index.

In general there are two types of file organization mechanisms that are used for indexing method conjunction to organize and store data:

1.Sequential File Organization also known as Ordered Index: In this case, the indices whichis based on the sorting of values according to their order of appearance. This is a classic form of storing mechanism that is generally fast, as it uses a physical layer to store information.The Sequential or Ordered file organization can be used to store the data into the sparse or dense way.:

(a) Dense Index:

In the data file, each search key value is indexed.

As well as the search key, this record contains a link to the very first data record containing the key value of search.

(b) Sparse Index:

The data file contains only few entries with index records.They all point to the same block.

The index record whose key value of search is equal or less than to the key value of search we require is the record we look for whenever we are looking for a record.

The index record is the first one we examine. From there, we follow the pointer which is available in the file and it should be sequentially until we locate the relevant record.

2. Hash File Organization: The index is a measure of the distribution of values over a set of buckets. The buckets to which values are assigned are determined by a function known as a hash function.

There are essentially three indexing methods:

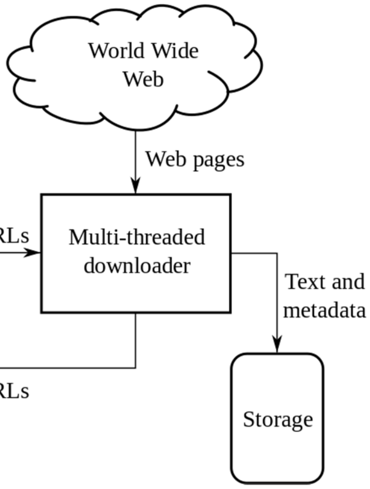
1. Indexing in Clusters
2. Secondary or non-clustered indexing
3. Indexing on several levels

Ranking

After compiling a list of hundreds of Web sites in response to a inquiry of user, a search engine will rank the results before they can be displayed to the user. A search engine is useless unless it can display the most applicable documents are shows at the top of the results page. Users must refine their queries by using advanced features of search, in addition to those mentioned previously, to minimize the number of unapplicable responses while maximizing those that are applicable, because the ranking algorithms frequently fail to place applicable documents at the top of a list of the responses.

The bulk of search engines use keywords to rank results. A simple ranking system would provide a higher ranking to a document containing all of the query's keywords and the lower ranking to one containing only a portion of them. In this simple formula, the keyword weights stored into the database of search engine are taken into account. The weights for the matched keywords can be added together to generate a score for a document. The document will be awarded a higher score if the matching keywords is one of the most important terms of the document. Each keyword that appears in the document which is query keyword is assigned numerical weights, which determine the score of the document.

IV. Working Methodology



As described in the preceding section, each crawler would have a robust crawling strategy and a well optimised design. According to Shkapenyuk and Suel, While creating crawler that can download few pages per second is quite simple but it is slow, Building a highperformance machine that could download for a short length of time Thousands of millions of pages spread out over some weeks causes a slew of data management challenges.

There are several elements to consider, including system architecture, network efficiency and input and output, as well as controllability and ruggedness. The algorithms and design of web crawlers are kept as trade secrets. There is usually a lack of detail in crawler designs when they are published, making it impossible for others to duplicate the work. Concern about the "search engine spamming," which deflect textensive search engines from releasing their algorithmic ranks, are also growing.

Working of Google Search Engine

Google collects information from a variety of sources and including online pages that are usersubmitted content (such as your ‘Google Business Profile’ and ‘user submissions’ to Google Maps), public internet databases, book scanning and a range of other sources.This page, on the other hand, is about web pages. To generate the results from web pages, Google uses 3 main steps:

Crawling

The first step is to look up internet pages. Google should constantly look for some new web pages for adding to its database considering there is no central database of all websites.

Because of previous visits of Google, some pages are well-known. If a known page links to a new one, Google learns about the new page. Google detects new pages by scanning a list of pages provided by a web site owner. In the event that you change or add to a site managed by a web hosting company, such as Blogger as well as Wix, they could inform Google about it.

When Google finds the URL of page, it visitsor crawls to see what is on it. The search results should show the pages where they belong, This image is rendered and analyzed by Google for its textual information and non-textual information, as well as its visual arrangement. Google will be able to more accurately match your content to those who are searching for content like yours when it better understands your site.

To boost the crawling of your website, do the following:

Google should be able to access the pages of your site and the content should look correct.

The Google search engine searches the web anonymously (one who doesn't have a username or password). information). All of the photos and other aspects of the website must be visible to Google. page in order to fully comprehend it.

DOI: 10.48175/568

You can submit a specific URL to if you've built or changed a single page.

Google. Use a sitemap to notify Google about a large number pages that are new as well as updated at once. Make your home page if you just want Google to crawl one page. Google considers this home page to be the most important page on this website. To encourage the thorough website crawl, make sure this home page has a batter site.

Users (and Google) will be able to navigate our site more easily if we have a navigation system this could links to every of the major pages or components. For small-scale sites, make Google wise tothethis homepage is adequate, as long as Google could hold out all of the other pages of this site via a chain of connections that starts with homepage of site.

Obtain a link to your page from another website that Google already knows about. Links in advertisements, links in comments,paid connections on another sites, links in comments, and other link that do not match the Guidelines of Google Webmasterthat will Google not follow.

Indexing

When a page is discovered by google it tries to find out what it's about. This method is known as indexing. The search engine analyses the text on the page, catalogues any images or videos it finds, and tries to comprehend the page as a whole.

Google maintains this information in its index, which is a vast database that spans many computers.

Your page indexing can be improved by doing the following:

Create page titles that are succinct and meaningful.

Use page titles that communicate the page's topic.

To convey information, use language rather than graphics. Google recognises some images and videos, but not as effectively as it recognises text. At the very least, include alt text and other appropriate characteristics to annotate your video and photographs.

Ranking

Google tries to find the best answers by assessing factors such as the location of user, language and device and factoring in additional aspects that will furnish the better user experience as well as most applicable answer. When searching for "bicycle repair shops," a user in Paris, for example, might get different results than a user in Hong Kong. Google does not accept repay in exchange for higher ranks, as they are determined by an algorithm.

To improve your ranking and serving, are doing the following:

Make your website mobile-friendly and load quickly.

Maintain your page by adding helpful content and keeping it up to date.

To ensure a great user experience, follow the Google Webmaster Guidelines.

V. Conclusion

L. PageandS. Brin focus on the fundamental notions of the Web crawler and present a Web crawling algorithm. PyBot which is the simple BFS based crawler which is crawled the collected data and university site to start the investigation. Furthermore, the crawler

I crawled a few additional sites and compared information like entire pages per site, dangling pages, non-dangling pages, external pages, crawling time and crawling pages. The PageRank algorithm was implemented using the Crawler, which is not detailed here because it is outside the purview of A. Gulli and A. Signorini. One future project could be to perform research in the subject of indexing and evaluation, then include these modules into PyBot crawler to build a fully-fledged Web Crawler.

REFERENCES

[1]. S. Amudha, “Web Crawler for Mining Web Data” in International Research Journal of Engineering and Technology Volume: 04 Issue: 02, Feb -2017

[2]. Ayar Pranav, Sandip Chauhan, “Efficient Focused Web Crawling Approach for Search Engine”, International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 5, May 2015

DOI: 10.48175/568

[3]. Mini Singh Ahuja, Dr Jatinder Singh Bal, Varnica, "Web Crawler: Extracting The Web Data", International Journal of Computer Trends and Technology (IJCTT) – volume 13 number 3 – Jul 2014

[4]. S. Brin, and L. Page, "The anatomy of a large-scale hypertextual Web search engine". Proc. of the WWW Conference (WWW'98), pp 107-117, 1998

[5]. Christopher Olston, Marc Najork, "Web Crawling", Foundations and Trends in Information Retrieval Vol. 4, No. 3 (2010) 175–246

[6]. Vandana Shrivastava, "A Methodical Study of Web Crawler", Vandana Shrivastava Journal of Engineering Research and Application, Vol. 8, Issue 11 (Part -I) Nov 2018