International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Building Your Own Search Engine

Mansi Vinzani[1], Divya Salian[2], Priyesh Bapna[3], Prof. Sneha Bendale[4]

[1,2,3]Student, Department of Computer Engineering, Terna Engineering College, Maharashtra, India
[4]Professor, Department of Computer Engineering, Terna Engineering College, Maharashtra,, India

---

Abstract – *21st Century being an age of digital world makes every information available online and to be able to access such huge amount of information on variety of subjects is only possible by having a search engine. We have so many search engines available by now but the most prominent search engine in the world right now is 'Google Search Engine'. Every other search engine is different from that of others. The reason for this difference is primarily the use of algorithms or the combination of algorithms in use. Every algorithm will have its own pros and cons. Our project titled as 'Building our own search engine' is based on the concept of web-crawling and indexing. The implementation of these 2 will form the major chunk of our project. Our web-crawler (spider/bot) will be essentially sorting through the Internet to find website addresses and the contents of a website for storage in the search engine database. Then comes the job of our Indexer which will be indexing the content based on the occurrence of keyword phrases in each individual website. Also, our search engine would store the web content within the database essential for fast and easy searching. Finally the output of our project are the hyperlinks to websites that show up in the search engine page when a certain keyword or phrase is queried.*

***Key Words*: Crawler, Heap sorting, Page Scoring, Breadth first search.**

1. INTRODUCTION

A search engine is a web-based tool which allows the internet users to find information on the internet. Most commonly used search engines are Google, Yahoo!, MSN, Bing, Ask etc. Search engines are special types of programs used to search documents having specified keywords and returns a list of documents where the keywords are located. A search engine is usually a general collection of programs. However, the term 'search engine' is often used to generally describe the common systems like Google, Bing and Yahoo! Search engines generally use automated software applications e.g. robots or spiders which moves across the Web and follows the links from page to page, site to site. A typical Web Search Engine starts working by sending out a spider which has the ability to fetch as many documents as possible against the supplied keywords. There is another program called the indexer, and then starts reading these documents and starts creating the indices based on the tokens found in each document. Each search engine uses its own proprietary algorithm to create the indices in such a way that ideally, only meaningful results are returned for each query. A typical search engine consists of few parts – crawler which is used to pull external documents -An index which is the place where the documents are stored in an inverted tree and a document store to keep the documents.

2. LITERATURE SURVEY

[1] A survey on Search Engine Optimization is done. Survey on different search engines are made to check how user friendly, stable, secure and fast the search engine is. Different optimization techniques are used such as Keyword Search, Web optimization, Social Media, Blogs, Forums, Articles, Analytics and Press Releases.

[2]Different parts of search engines are explained. Basic Features of search engine is explained. Types of search engines such as Search directories or indexes, Hybrid search engines and Meta search engine work is explained. There are three important parts of search engine i.e. Spider or Crawler, Index, catalog or database and Search Engine Software.

[3]Major Search Engines like Google, Yahoo and Bing are being compared. Different techniques and comparisons are made. Speed, Accuracy, Security, Related Content, and no of results which are useful are compared. Google and Yahoo turns out to be the best but when it comes to no of useful results, Google turns out to be the best.

[4]A search engine to support web terrorism mining offers un identifies and precious data for the Law enforcement organization incessantly. The process to tackle examining the criminal believes of forensic data investigation regarding prioritizes capable links and pages which wrap the reliability and consistency gap by offering a framework. This argue with variety of open issues in this area.

[5]Bee Swarm algorithm is used instead of regular breadth first search algorithm in making of this search engine. Swarm intelligence is used to crawl the web instead of breadth first search. It works like a bee system to collect search results fast as well as accurate.

[6]Web mining and data mining techniques is used to make a personalized web search engine. Data overload problem can be overcome using web mining and data mining techniques. Search can be made Classifies as well as Personalized. Web personalization is done. Strategies of web personalization and their related work is made.

3. METHODOLOGY

Methodology is divided into four parts as follows:

3.1. Admin Process

Admin Login: Admin login into the system using username and password. After successful login, admin can access the crawler.

Dashboard: On this window search bar for query is available and there are two more buttons available i.e. 'Extract top 20 links' and 'Extract Max URLS'. Extract max URLs will show the URL with the highest page score.

3.2. User Query

User is provided by a search bar in which user types the keyword that he/she wants to have information about. The result will be displayed to the user in the form of URL links. First 30 links on the basis of page score will be displayed.

3.3. Web Crawler

The web crawler begins and works independently. To initiate the crawler we need to provide the seed URL links to it. Tracing which it can independently download the websites. The web crawler goes from one link to another in a breadth first search manner so that all primary relevant sites are crawled first.

3.4. Indexer

After the web crawler has crawled the websites, it will be saving it into a database. An online database is used for this purpose.

3.5. Page Scoring

Page Scoring is done with the help of calculating 2 scores i.e. Keyword score and link score. Keyword score is on the basis of occurrence of that keyword on the webpage as to how many times that was word has been encountered in the website. Word count of that keyword is recorded. Second one is by calculating Link score. Link score is determined by tracing all the href tags on the webpage to get the count of outgoing links. This is also recorded and kept. Now the page score is calculated by adding the above two scores.

3.6. Searching Interface Steps :

Step 1: Start

Step 2: Admin Login

Step 3: Enter the keyword for your search.

Step 4: Hit the search button

Step 5: Result displays 30 URLs

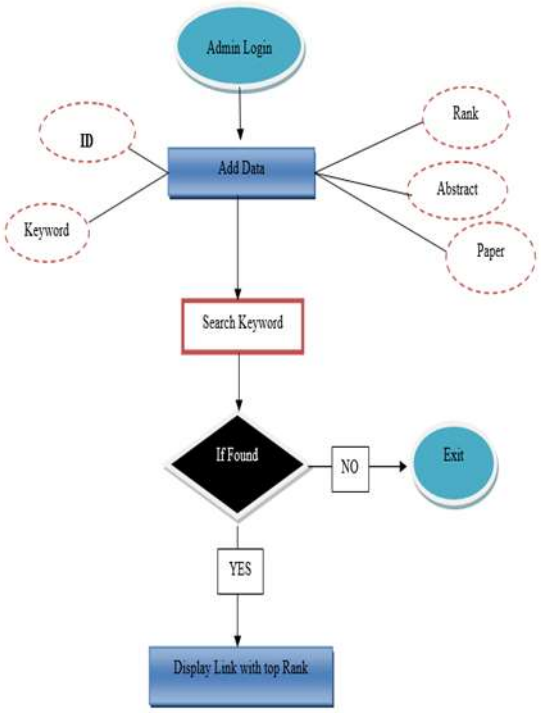
Step 6: Hit the button 'Top 20 links"

Step 7: Result displays top 20 links on the basis of page score

Step 8: Hit the button 'Extract max URLs'

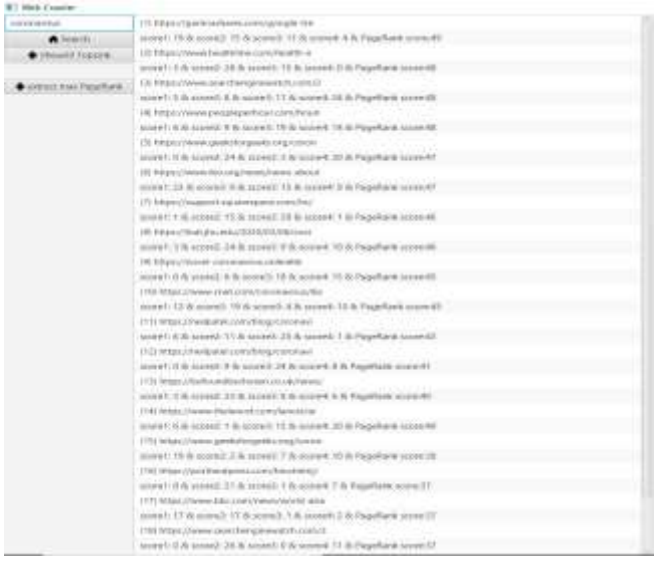
Step 9: Result will display one such link which has the highest page score.

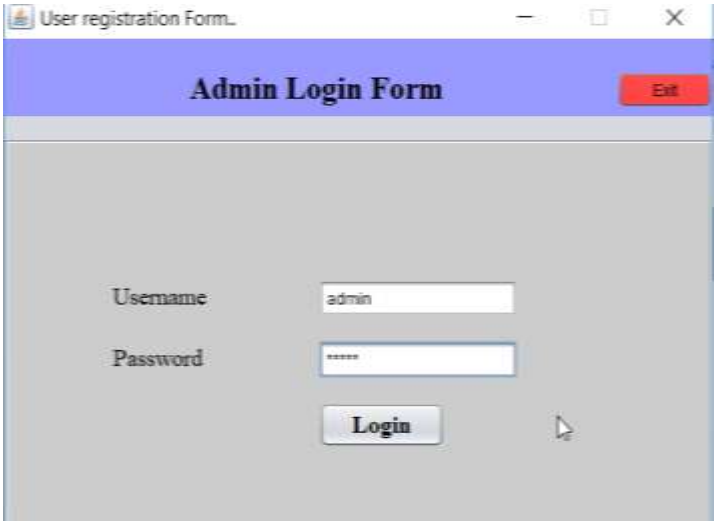
4. FLOW CHART OF THE SYSTEM



4.1. Flowchart

5. RESULTS





6. CONCLUSION

Search Engine which is one of the most powerful tool crawls up so many websites to bring up the result on the display screen. Every search engine will have different results being displayed on the top because of the primary reason of using different algorithms to crawl and ranking the pages. Further by optimizing the search algorithm more relevant websites can be displayed on the top, increasing the efficiency of the search.

7. REFERENCES

1] Prashant Ankalkoti "Survey on Search Engine Optimization", Imperial Journal of Interdisciplinary research, vol – 3, issue – 5, 2017.

[2] W.Bruce croft, Donald Metzler, Trevor Strohman " Search Engines", Pearson Education Inc, 2015.

[3] Bhupendra Ratha "Search Engine", School of Library and Information Science.

[4] Amim SHahraki, Javad Hosseinkhan, Surayti Chuprat "Applying Social Network Analysis in Crawler Based Search Engine", International Journal of Computer Science and Network Security, VOL. 17No. 8, August 2017.

[5] Mohammed Ibrahim Sahujaa, Ahmed bahaa Ulddin "Building Web Crawler Based on Bee Swan Algorithm", International Journal of Computer Science and Network Secrity, Vol. 10, Issue 5, No, 1, September 2013.

[6] K. Ppoornsiri, S.Radha Priya "A Literature Review on Personalized Web Search", Research Department of Computer Science, Coimbatore, India.