International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Building Search Engine using Machine Learning

Vani Chinni, Shravika Challa, Chakradhar Annandi

CH.VANI (CSE,CMR Technical Campus)
CH.SHRAVIKA(CSE,CMR Technical Campus)
A.CHAKRADHAR(CSE,CMR Technical Campus)

Abstract –

The web is the huge and most extravagant wellspring of data. To recover the information from the World Wide Web, Search Engines are commonly utilized. Search engines provide a simple interface for searching for user query and displaying results in the form of the web address of the relevant web page, but using traditional search engines has become very challenging to obtain suitable information. This paper proposed a search engine using Machine Learning technique that will give more relevant web pages at top for user queries.

Key Words: web crawling, Indexer, Search Engine,Webpages.

1.INTRODUCTION

World Wide Web is actually a web of individual systems and servers which are connected with different technology and methods. Every site comprises the heaps of site pages that are being made and sent on the server. So if a user needs something, then he or she needs to type a keyword. Keyword is a set of words extracted from user search input. Search input given by a user may be syntactically incorrect. Here comes the actual need for search engines. Search engines provide you a simple interface to search user queries and display the results.

1) Web crawler : Web crawlers help in collecting data about a website and the links related to them. We are only using web crawlers for collecting data and information from WWW and storing it in our database.

2) Indexer : Indexer which arranges each term on each web page and stores the subsequent list of terms in a tremendous repository.

3) Query Engine : It is mainly used to reply to the user's keyword and show the effective outcome for their keyword. In the query engine, the Page ranking algorithm ranks the URL by using different algorithms in the query engine.

4)This paper utilizes Machine Learning Techniques to discover the utmost suitable web address for the given keyword. The output of the PageRank algorithm is given as input to the machine learning algorithm.

2. Body of Paper

The Crawler based Search Engines Crawler based search engines[3] such as Google create their listings automatically. They crawl or spider the web, then people search through what they have found. If you change your webpages, crawler based search engine will find these changes and that can affect how you are listed. Three elements in crawler based search engines are:

• Crawler or spider

• index or catalog

• search engine

software Crawler or spider visits webpages and reads it and index or catalog is like a giant book containing a copy of every webpage that crawler or spider finds. If a webpage changes, then this book is updated with a new one. Human Powered Directories A human powered directory such as the open directory depends on humans for its listings. In this type of search engine, site owner submits a short description of the site to the directory along with category it is to be listed. Submitted site is then manually reviewed and added in the appropriate category or rejected for listing. Keywords entered in a search box will be matched with the description of the sites. This means the changes made to the content of web pages are not taken into consideration as it is only the description that matters. A good site with good content is more likely to be reviewed for free compared to a site with poor content.

Meta Search Engines Meta search engines[4] gives results based on a combination of results from other search engine databases. It uses complex algorithms and virtual databases. A search engine that queries other search engines and then combines the results that are received from all. In effect, the user is not using just one search engine but a combination of many search engines at once to optimize web searching. For example, Dog pile is a meta search engine directories as secondary mechanism. For example, google may take the description of a webpage from human powered directories and show in the search results. As human powered directories are disappearing, hybrid types are becoming more and more crawler based search engines.

But still there are manual filtering of search result happens to remove the copied and spammy sites. When a site is being identified for spammy activities, the website owner needs to take corrective action and resubmit the site to search engines. The experts do manual review of the submitted site before including it again in the search results. In this manner though the crawlers control the processes, the control is manual to monitor and show the search results naturally. B.5 Speciality Search Engines Speciality search engines search a specially created database which is limited to a particular subject. A

speciality search engine, sometimes called a topican or vertical search engine, searches a specially-created database limited to a particular subject. Speciality search engines fall into two main categories: • service • subject-specific Speciality service search engines provide services that are often not available from larger general search engines. Subject-specific search engines search a database tailored to a particular subject. Depending on your area of interest and the type of information you are seeking, speciality search engines can provide more relevant results more quickly than a general purpose search engine such as Google or Yahoo.

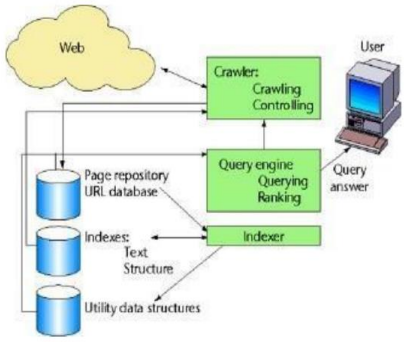

Fig-1: System Architecture

Because of this it would be wise to also submit your blog or website to some of the speciality search engines that cater for your niche. C Search Engine Working While you should always create website content geared to your customers rather than search engines, it is important to understand how a search engine works. Most search engines build an index based on crawling, which is the process through which engines like Google, Yahoo and others find new pages to index. Mechanisms known as bots or spiders crawl the web looking for new pages. The bots typically start with a list of website. URL's determined from previous crawls. When they detects new links on these pages, through tags like HREF and SRC, they add theses to the list of sites to index. Then, search engine use their algorithms to provide you with a ranked list from their index of what pages you should be most interested in based on the search terms you used.

Then, the engine will return a list of web results ranked using its specific algorithm. On Google, other elements like personalized and universal results may also change your page ranking. In personalized results, the search engine utilizes additional information it knows about the user to return results that are directly catered to their interests. Universal search results combine video, images and Google news to create a bigger picture result, which can mean greater competition from other websites for the same keywords. Search engine optimization is a set of rules that can be followed by website owners to optimize their websites for search engines and thus improve their search engine ranking. In addition, it is a great way to increase the quality of your website by making it user friendly, faster and easier to navigate. Steps in search engine optimization are as follows: • Website analysis • Client requirements

• Keyword research

• Content writing

• Website optimization

3. CONCLUSION

Web page ranking is a global ranking of all web pages, regardless of their content, based solely on their location in the web's graph structure. Using these web page ranking techniques, we are able to order search results so that more important and central web pages are given preference. All things considered, search engine optimization will become more resourceful in the upcoming years, but also more complex, forcing marketers to develop more elaborate strategies that bring more types of content, devices and tools into the equation. But no matter which combination of elements you see, the focus should stay on the user and their needs, as machine learning and artificial intelligence technologies will transform ranking factors that can better reflect the needs and expectations of searchers.

ACKNOWLEDGEMENT

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project. We take this opportunity to express my profound gratitude and deep regard to my guide.

Dr. G. Madhukar, Associate Professor for his exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing, help and guidance given by him shall carry us a long way in the journey of life on which we are about to embark. We also take this opportunity to express a deep sense of gratitude to Project Review Committee (PRC) **Dr. M. Varaprasad Rao**, **Dr. G.Madhukar**, **Dr. T. S. Mastan Rao**, **Dr. Suwarna Gothane**, **Mr. A. Uday Kiran**, **Mr. A. Kiran Kumar**, **Mrs. G. Latha** for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to **Dr. K. Srujan Raju**, Head, Department of Computer Science and Engineering for providing encouragement and support for completing this project successfully.

We are obliged **to Dr. A. Raji Reddy**, Director for being cooperative throughout the course of this project

We also express our sincere gratitude to **Sri. Ch. Gopal Reddy**, Chairman for providing excellent infrastructure and a nice atmosphere throughout the course of this project.

The guidance and support received from all the members of **CMR Technical Campus** who contributed to the completion of the project. We are grateful for their constant support and help. Finally, we would like to take this opportunity to thank our family for their constant encouragement, without which this assignment would not be completed. We sincerely

acknowledge and thank all those who gave support directly and indirectly in the completion of this project.

REFERENCES

1. Vishwas Ravall and Padam Kumar, "SEReleC (Search Engine Result Refinement and Classification) – A meta search engine based on combinatorial search and search keyword based link classification," in IEEE-International Conference on advances in Engineering, science and management(ICAESM-2012), March 30,31,2012.Saad ALBAWI, Tareq Abed MOHAMMED, Saad AL-ZAWI, "Understanding of a convolutional neural network," ICET 2017.

2. Vijay Chauhan, Arunima Jaiswal, Junaid Khalid khan, "Web page ranking using machine learning approach," in Fifth International Conference on Advanced Computing and Communication Technologies2015.T. Yamunarani, G.Kanimozhi, "Hand gesture recognition system for disabled people using arduino," vol. 4, 2018

3. Farzaneh Shoeleh, Mohammad Sadegh Zahedi, MoiganFarhoodi, "Search Engine Pictures: Empirical analysis of a web search engine query log," in Third International Conference on Web Research(ICWR), 19,20 April 2017.

4. Donghong Liu, Xan Xu, Yu Long, "On member search engine selection using artificial neural network(ann) in meta search engine," in 2017.

5. P. Kshirsagar, S. Akojwar, Nidhi D. Bajaj , "A hybridised neural network and optimisation algorithms for prediction and classification of neurological disorders" International Journal of Biomedical Engineering and Technology ,vol. 28,Issue 4,Pp. 307-321,2018.