IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SEARCH ENGINE USING WEB CRAWLING

D.Sujeetha.M.E[1],Bhuvaneswari.D[2],Devi Priya.K[3],Indhumathi.M[4]

[1]Assistant Professor,[2,3,4] UG Students

Department of Computer Science and Engineering,
Mahendra Engineering College, Namakkal, India-637503

Abstract: Today's world the information is most valuable quantity. With the advent of the web, the information storage and retrieval have taken a huge step forward. Search engines plays important role in this area. In this report (implementation of a search engine), we talk about the functionality of a mini-offline search engine. We study the various components of search engine is which involves "crawlers" (a spider program to search through documents), "porter and stemmer" (program that remove stop words and brings the query in its basic form) and "indexer" (one which indexes the documents to cut short the duration of searching). Next part is the implementation of these various components .The search engine while searching through the web gives us so many relevant and irrelevant. Which relevant information should come as a best desired result, depends on the kind of algorithms that all of the search engine have got the propriety right over. I have also tried to develop a similar algorithms named as page rank, which has been implemented on a small web graph and extract the information through the in-links and out-links of any web page.

Keywords: Search Engine, Page Rank, Web Mining, Links, World Wide Web,web Crawler.

I. INTRODUCTION

Many papers written in the web information retrieval (IR) field utilize their own web crawlers to crawl, index, and analyze contents (including hyperlink texts) of the pages and network structure of the web. Sometimes a search function to return relevant pages to the users' queries is also provided. Crawler and search function are considered to be the fundamental components of a search engine [1], and each has its own research challenges and problems.

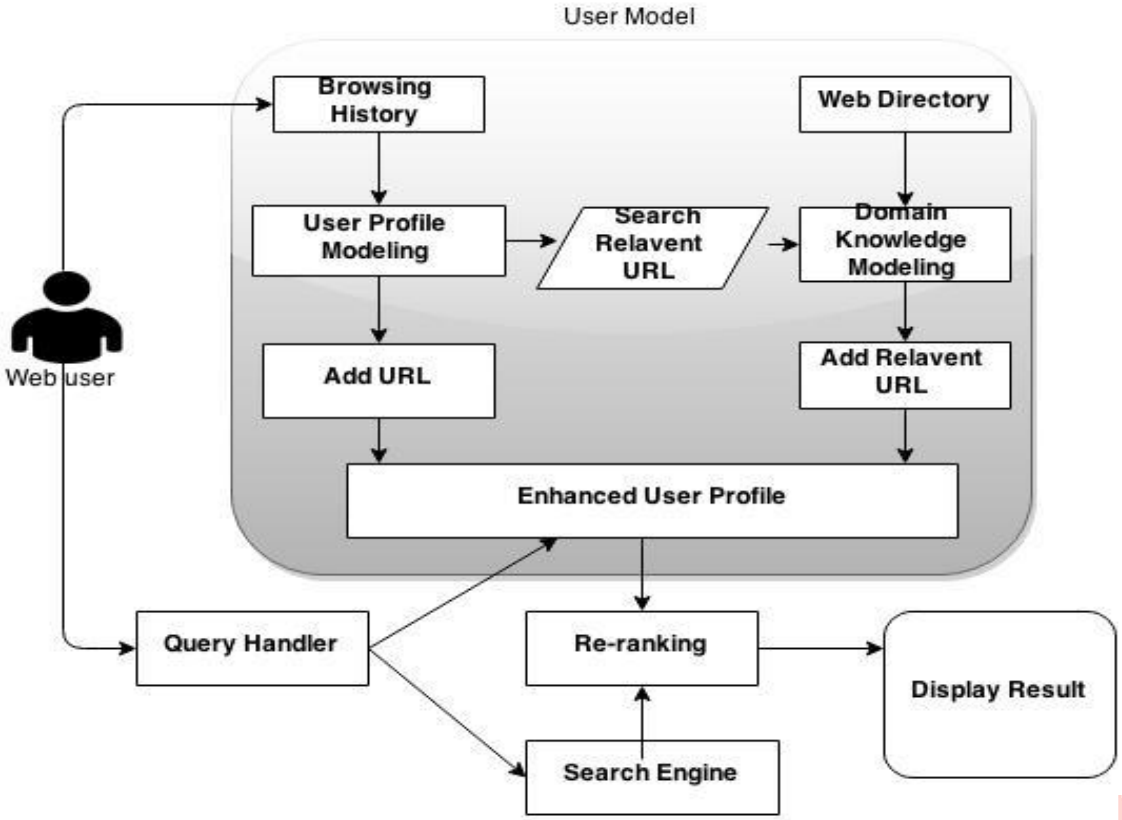
Web crawler, also known as spider or robot, is responsible for fetching pages, parsing hyperlinks, managing crawl queue, and indexing contents of the pages. In a more sophisticated form, this component also implements politeness policies (e.g., obeying robot.txt instructions and not overloading servers with repetitive pages queries [1, 2, 3]), indexes anchor text [4], and decides in advanced which are spam links or pages with different URLs but similar contents (so that the crawler can avoid downloading these pages to save bandwidth [1, 5, 6]). Some of the real working crawlers like IRLbot [1], Mercator [2, 7], Polybot [3], iRobot [6], UbiCrawler [8], and Googlebot [9] provide good documentations on their designs and implementations. However, the descriptions are still too general to make any reproduction effort possible.

Search function is an interface between a search engine and users. This function receives users' queries and returns relevant pages to the queries. The pages usually are sorted according to some criteria. The most basic criterion is Boolean match between contents of the pages and words in the queries. More advanced mechanisms use hyperlink structure of the web graph (e.g., PageRank [9], HITS [10], and Salsa [11]), anchor text information [12, 13, 14, 15], and user-click behavior [4, 16] to determine ranking of the relevant pages.

There are some notable open source search engine projects like Lemur (lemurproject.org) and Lucene (lucene.apache.org). But due to their complex design and implementation, these projects are still not a good starting point to learn about web search engine technology. Here the importance of open source search engine project that the design and implementation are easy to understand is emphasized because there are many occasions that we want to do.

2.SYSTEM OVERVIEW

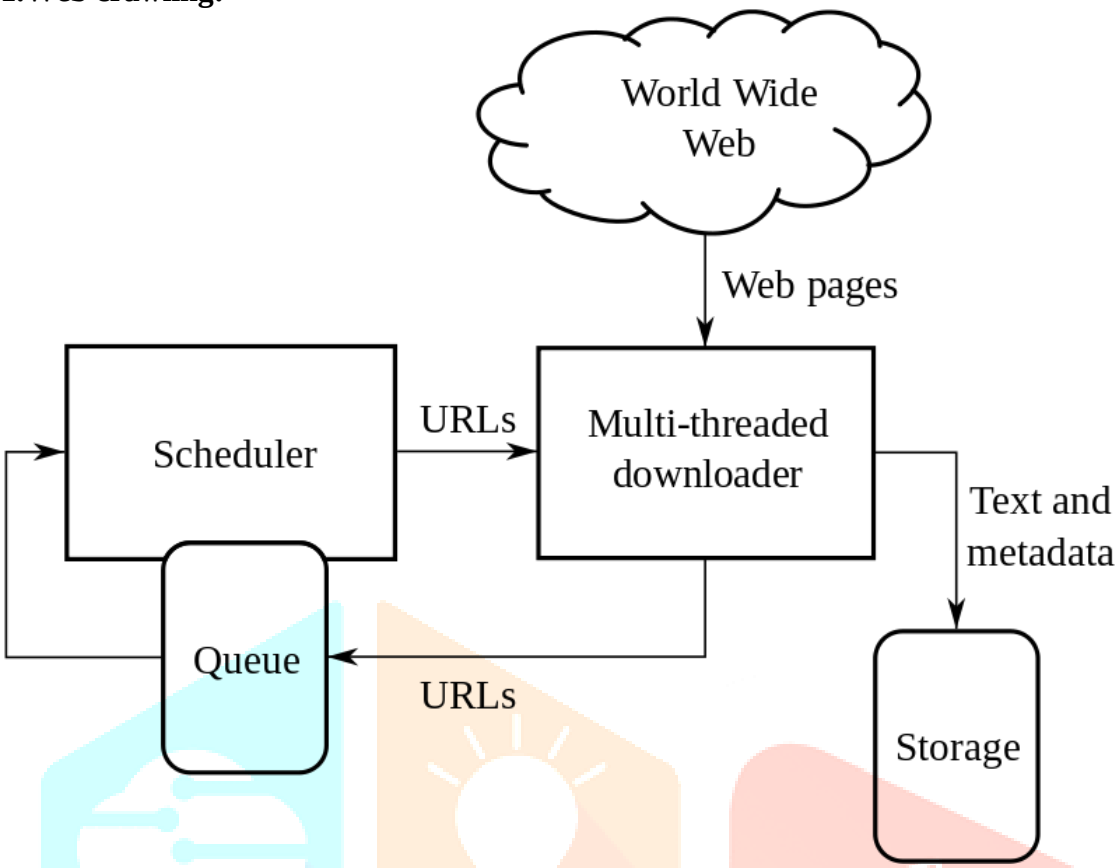
The system is designed and implemented by pertaining the original architecture. In the top level view, the system consists of crawler, searcher, and database components.



3.HOW DO SERACH ENGINES WORK

1. Web crawling
2. Indexing
3. Searching

1.Web crawling:



A **Web crawler**, sometimes called a **spider** or **spiderbot** and often shortened to **crawler**, is an Internet bot that systematically browses the World Wide Web, typically operated by search engines for the purpose of Web indexing (*web spidering*).

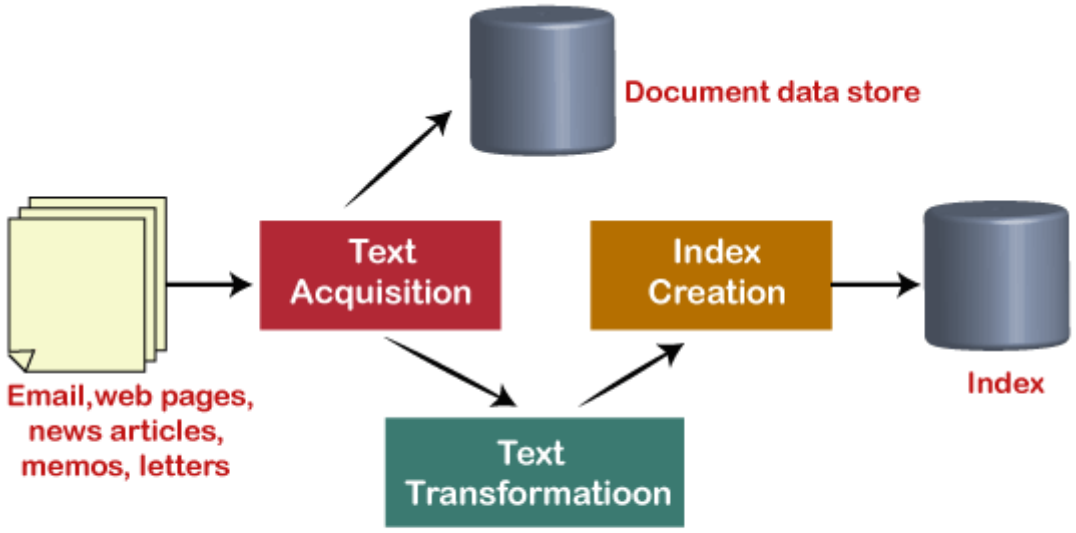
Web search engines and some other websites use Web crawling or spidering software to update their web content or indices of other sites' web content. Web crawlers copy pages for processing by a search engine, which indexes the downloaded pages so that users can search more efficiently.

Crawlers consume resources on visited systems and often visit sites without approval. Issues of schedule, load, and "politeness" come into play when large collections of pages are accessed. Mechanisms exist for public sites not wishing to be crawled to make this known to the crawling agent. For example, including a robots.txt file can request bots to index only parts of a website, or nothing at all.

The number of Internet pages is extremely large; even the largest crawlers fall short of making a complete index. For this reason, search engines struggled to give relevant search results in the early years of the World Wide Web, before 2000. Today, relevant results are given almost instantly.

Crawlers can validate hyperlinks and HTML code. They can also be used for web scraping and data-driven programming.

2.Indexing



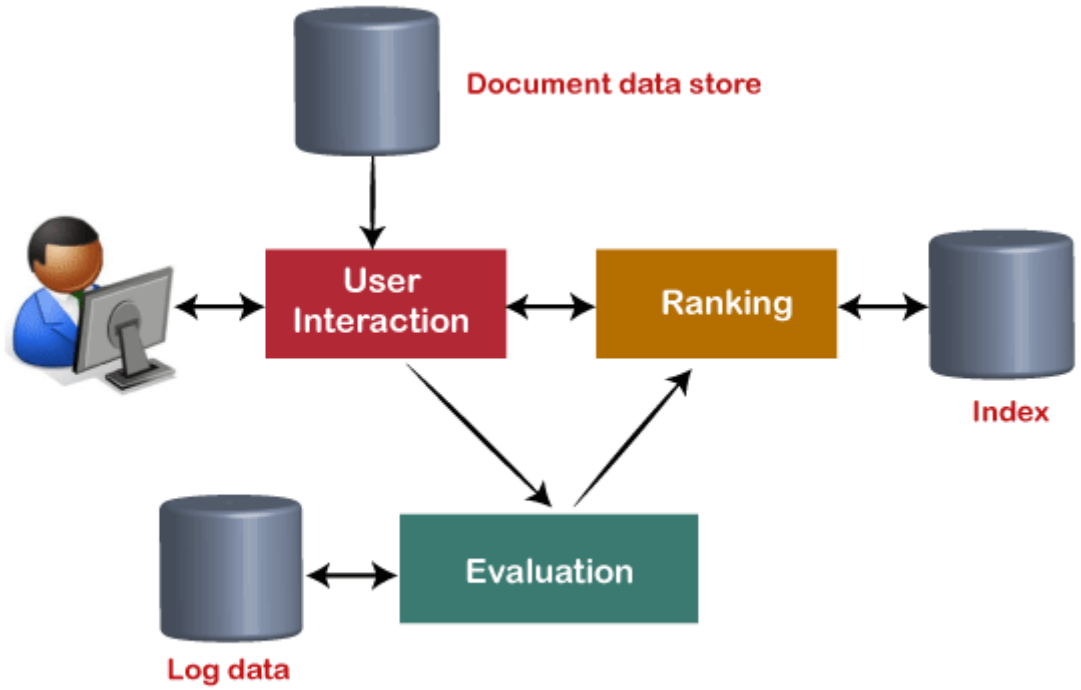
The indexing process

Search engine optimisation indexing is the collecting, parsing, and storing of data to facilitate fast and accurate information retrieval. Index design incorporates interdisciplinary concepts from linguistics, cognitive psychology, mathematics, informatics, and computer science. An alternate name for the process in the context of search engines designed to find web pages on the Internet is web indexing.

Popular engines focus on the full-text indexing of online, natural language documents. Media types such as video, audio, and graphics are also searchable.

Meta search engines reuse the indices of other services and do not store a local index, whereas cache-based search engines permanently store the index along with the corpus. Unlike full-text indices, partial-text services restrict the depth indexed to reduce index size. Larger services typically perform indexing at a predetermined time interval due to the required time and processing costs, while agent-based search engines index in real time.

3.Searching



The query process

A **web search query** is a query based on a specific **search term** that a user enters into a web search engine to satisfy their information needs. Web search queries are distinctive in that they are often plain text or hypertext with optional search-directives (such as "and"/"or" with "-" to exclude). They vary greatly from standard query languages, which are governed by strict syntax rules as command languages with keyword or positional parameters.

4.Use of Search Engine in Information Retrieval

Information retrieval is the process of satisfying user information needs that are expressed as textual queries. Search engines represent a Web-specific example of the information retrieval paradigm. The problem of Web search has many additional challenges, such as the collection of Web resources, the organization of these resources, and the use of hyperlinks to aid the search. Whereas traditional information retrieval only uses the content of documents to retrieve results of queries, the Web requires stronger mechanisms for quality control because of its open nature. Furthermore, Web documents contain significant meta-information and zoned text, such as title, author, or anchor text, which can be leveraged to improve retrieval accuracy.

5.Web mining

Web Mining is the process of Data Mining techniques to automatically discover and extract information from Web documents and services. The main purpose of web mining is discovering useful information from the World-Wide Web and its usage patterns.

Applications of Web Mining:

1. Web mining helps to improve the power of web search engine by classifying the web documents and identifying the web pages.
2. It is used for Web Searching e.g., Google, Yahoo etc and Vertical Searching e.g., FatLens, Become etc.
3. Web mining is used to predict user behavior.
4. Web mining is very useful of a particular Website and e-service e.g., landing page optimization.

Web mining can be broadly divided into three different types of techniques of mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. These are explained as following below.

1. Web Content Mining:

Web content mining is the application of extracting useful information from the content of the web documents. Web content consist of several types of data – text, image, audio, video etc. Content data is the group of facts that a web page is designed. It can provide effective and interesting patterns about user needs. Text documents are related to text mining, machine learning and natural language processing. This mining is also known as text mining. This type of mining performs scanning and mining of the text, images and groups of web pages according to the content of the input.

2. Web Structure Mining:

Web structure mining is the application of discovering structure information from the web. The structure of the web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Structure mining basically shows the structured summary of a particular website. It identifies relationship between web pages linked by information or direct link connection. To determine the connection between two commercial websites, Web structure mining can be very useful.

3.Web Usage Mining:

Web usage mining is the application of identifying or discovering interesting usage patterns from large data sets. And these patterns enable you to understand the user behaviors or something like that. In web usage mining, user access data on the web and collect data in form of logs. So, Web usage mining is also called log mining.

6.Page Rank

every **Search Engine** uses its own algorithm to **rank** webpages making sure that only relevant results are returned for the query entered by the user. The result for a specific query is then shown on the **Search Engine** Results **Page** (SERP). **Rankings** in SEO refers to a website's position in the **search engine** results page. There are various **ranking** factors that influence whether a website appears higher on the SERP based on the content relevance to the **search** term, or the quality of backlinks pointing to the page.

7.Algorithms

A search engine spider, also known as a **web** crawler, is an Internet bot that crawls websites and stores information for the search engine to index. When you search something on Google, those pages and pages of results can't just materialize out of thin air.

8.Conclusion

This report introduced the functionality of a Search Engine. our implementation of the parts of Search Engine have been developed and coded. We have used the Linux as part of implementation. The page rank introduces the concepts of giving importance to the web pages that a search engine has to provide. Performance of this page rank algorithm was tested on web pages and has the capacity to be improved further with sparse matrix concepts.

9.References

T. Segaran, Programming Collective Intelligence: Building Smart Web 2.0 Applications, O'Reilly Media Inc. (2019) pp. 49—52.

Z. Bar-Yossef, I. Keidar, and U. Schonfeld, ―Do Not Crawl in the DUST: different URLs with similar text,‖ in Proc. 16th International WWW Conference (2019) pp. 111—120.

R. Cai, J.M. Yang, W. Lai, Y. Wang, and L. Zhang, ―IRobot: An Intelligent Crawler for Web Forums,‖ in Proc. 17th International WWW Conference (2020) pp. 447—456.

A. Fujii, ―Modeling Anchor Text and Classifying Queries to Enhance Web Document Retrieval,‖ in Proc. 17th International World Wide Web Conference, (2019) pp. 337—346.

R. Baeza-Yates, L. Calderon-Benavides, and C. Gonzales-Caro, ―The Intention Behind Web Queries,‖ in Proc. 13th International Conference on String Processing and Information Retrieval, (2019) pp. 98—109.

A. Mirzal and M. Furukawa, ―A method for accelerating the HITS algorithm‖, Journal of Advanced Computational Intelligence and Intelligent Informatics, (2018) Vol. 14, No. 1, pp.89–98.

K.S. Jones, S. Walker, and S.E. Robertson, ―A Probabilistic Model of Information Retrieval: Development and Comparative Experiments,‖ Information Processing and Management Vol. 36, (2019) pp. 779—840.

N. Eiron and K.S. McKurley, ―Analysis of Anchor Text for Web Search,‖ in Proc. ACM SIGIR, (2018) pp. 459—460.

A.N. Langville and C.D. Meyer, Google's PageRank and Beyond: The Science of Search Engine Rankings, Princeton University Press (2019)

A. Mirzal, ―PyThinSearch: A Simple Web Search Engine,‖ Proc. Int. Conf. on Complex, Intelligent and Software Intensive Systems, IEEE Comp. Soc. (2019) pp. 1—8.