**PROJECT DESCRIPTION**

**Introduction:**

World Wide Web has become a vast resource for information, which has led to the problem of management of information on the web and satisfy the user need. It has become difficult to trace the information on the web which has led to the need of development of some information retrieval technique which can find, extract, filter and obtain the output required by the user. Most known tool used presently is a search engine like google, yahoo or Bing for information retrieval. These tools answers millions of queries every day and index and rank each web page hundreds of millions of webpages. User inputs his query in the search engine and then search engine fetches web pages which are desired by the user from world wide web.

Query

USER

World Wide Web

Results

A search engine has four core components that are crawler, indexer, page ranking algorithm and query serving. Crawler also known as spider has the function to traverses through the entire web and store information about each webpage. The stored web pages are then parsed to an indexer which builds the index based on upon the keywords present in the pages. After the web pages which have keywords similar to that of the query are obtained then before representing the web pages to the user, a mechanism is needed which only shows those pages to the user which are desired by the user. That is web pages which are more relevant to the user are kept on the top and those which may not be relevant at the bottom. So, for to do this a page ranking algorithm is needed which can present the user with desired results. Then finally query serving is needed which can take the input of the user and perform all the steps above.

In this paper, an actual real-time web based search engine is developed depending on the user needs it will have the function and further after development of the search engine different page ranking algorithm are tested on it to find the comparison between the algorithm and which algorithm can provide the best efficiency and user desired results.

Finally, the performance of the search engine is seen based on response time, query throughput and indexing speed

SURVEY:

1. Page Ranking Algorithms in Online Digital Libraries: A Survey

1Sumita Gupta, 2Neelam Duhan, 3Poonam Bansal, 4 Jigyasa Sidhu

This paper presents with different ranking algorithms for academic digital libraries and highlights comparisons of theses algorithms in context of performance. It points out to the situations where still even in search engine like Google where a search query for some scientific literature, book results in tutorials, news and article etc. So, to overcome this problem, digital libraries have been introduced to make retrieval mechanism efficient. The paper surveys page ranking algorithm and do extensive comparison study based on many parameters like main technique used, methodology, quality of results, complexity etc. The page ranking algorithm discussed are as follows: PAGERANK Algorithm, POPULARITY AND SIMILARITY BASED PAGE RANK Algorithm (PSPR), SIMRANK: PAGE RANK Algorithm, PAGE RANKING USING ANNOTATION BASED ON LANGUAGE MODEL Algorithm.

**References:**

1)Course Textbook: Search Engines: Information Retrieval in Practice, by Croft, Metzler and Strohman.

2)Wikipedia: For data dump file and for information.

3)Sumita Gupta, Neelam Duhan(Page Ranking Algorithms in Online Digital Libraries: A Survey)

4)Neelam Duhan, A. K. Sharma(Page Ranking Algorithms: A Survey)

Query

Result

Webpages

USER

WWW

Query Indexer

Ranking Algorithm

Query Processor

Indexer

Crawler

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm  Measures | PageRank | Weighted PageRank | SimRank |
| Technique | Web Structure Mining | Web Structure Mining | Web Content Mining |
| Concept | Web links structure, score on basis of backlinks | Web links structure, score on basis of backlink and forward link | Scored according to similarity between query and pages |
| Input | Incoming links | Incoming links and outgoing links | Document and query contents |
| Complexity | O(logN) where N denotes the web pages | <O(logN) where N denotes the web pages | O(N2) where N denotes the web pages |
| Relevancy | Less | Less (Higher than PageRank) | More relevant than PR and WPR |
| Quality of Results | Medium | Medium (Higher than PageRank) | High accuracy and efficiency |
| Strengths | Simple Algorithm, analyze whole link structure of web | Considers weight and provide with more efficient results | Effective analyze of pages on basis of little contextual information |
| Weakness | Results at time of indexing and not at query time | Relevancy ignored, computes score at single level | Efficiency effected by crawler efficiency. |