ABSTRACT

The major issues in web crawling are doing an incomplete search through a part of the hypertext, crawling based on back-links and retrieving the pages based on either link, page content or category. In crawling, it is not practically feasible to visit all the millions of web pages given by the search engine to find the required information. In spite of many research contributions available in web crawling for the retrieval of pages, there is a need to provide effective retrieval of relevant pages at earlier stages of crawling.

The primary objective of this research is to develop enhanced web crawling techniques for effective information retrieval based on focused crawling and to minimize the search results, by providing relevant seed URLs, reducing the number of crawling levels, skewing the search to the relevant pages and retrieve the pages effectively. Some measures that are encountered in the deployment of proposed crawling techniques are link-based measure, text content using Levenshtein distance, logarithmic measure, probability measure, title of the document, page content and snippet-based retrieval.

A new Logarithmic and Probability Measure Algorithm (LPMA) for web crawling is introduced to minimize the number of crawling levels. The algorithm uses link, text content using Levenshtein distance, logarithmic measure and probability measure to provide more number of relevant pages. The advantage of the method is that the probability measure makes use of the similar and also the dissimilar keywords for improving the effectiveness of the algorithm. Different anchor texts are given to retrieve the web pages and the effectiveness of the algorithm is evaluated by comparing with the existing Integrated Page Rank Algorithm and the Improved Weight Algorithm.
A Topic-Specific Algorithm (TSA) is explored based on user interests to skew the search to the relevant pages. The algorithm uses the link, text content using Levenshtein distance and probability method to fetch more number of relevant pages based on the topic specified by the user. Evaluation is made on different topics and for different threshold values and it is shown that more number of relevant pages is retrieved. The algorithm reduces the crawling levels and fetches the best web pages under user interests during the earlier period of crawling and is evaluated by comparing with the existing Improved Weight Algorithm.

An Integrated Four Measures Algorithm (IFMA) is developed for the relevant retrieval of pages. The method efficiently utilizes the link-based measure, logarithmic distance measure, text content similarity using Levenshtein distance and probability measure to find the relevancy of the web pages. An Integrated study has been performed on the combined four measures and the four measures individually for different anchor texts. The performance of the proposed crawler is evaluated and it outperforms the four individual measures and the Improved Weight Algorithm.

A novel Two-Level Approach (TLA) technique is deployed using two-levels to fetch more number of most relevant pages effectively at earlier stage of crawling. In the first level, the anchor texts are verified with the title of the document, the snippet and the URL path. If any two or all the three possibilities contains the anchor texts, then the URLs are considered as relevant URLs and are given to second level. In the second level, the page content of the document is verified with the frequency of the anchor texts. More numbers of relevant pages are retrieved at first level, since the irrelevant pages are discarded in the initial and subsequent crawling and skew the search to more relevant pages. After second level, the most relevant pages are retrieved effectively. This approach is compared with the existing Improved Weight Algorithm to test the efficiency.
The web crawling techniques suggested in this research offer a better solution to minimize the search results by providing relevant seed URLs, to reduce the number of crawling levels, to skew the search to the relevant pages and retrieve the relevant pages effectively. This research mainly ensures the enhancement of the web crawling techniques and to effectively retrieve more number of relevant pages in minimum crawling.