ABSTRACT

World Wide Web (WWW) is a huge repository of interlinked hypertext documents known as Web pages. Users access these hypertext documents via Internet. Since its inception in 1990, WWW has become many folds in size, now it contains more than 50 billion publicly accessible web documents distributed all over the world on thousands of web servers and still growing at exponential rate. It is very difficult to search information from such a huge collection of World Wide Web as the web pages/documents are not organized as books on shelves in a library, nor are web pages completely catalogued at one central location. Search engine is basic information retrieval tool, used to access information from WWW. Users provide search queries in the Search engine’s interface. In response to the search query provided, Search engines use their database to search the relevant documents and produce the result after ranking on the basis of relevance. In fact, the Search engine builds its database, with the help of Web Crawlers, where a WebCrawler is a program that traverses the Web and collects information about web documents. To maximize the download rate and to retrieve the whole or significant portion of the Web search engines run multiple crawlers in parallel.

Overlapping of downloaded web documents, quality, network bandwidth and refreshing of web documents are the major challenging problems faced by existing parallel web crawlers that are addressed in this work. A Multi Threaded (MT) Server based novel architecture for incremental parallel web crawler has been designed that helps to reduce overlapping, quality and network bandwidth problems. Additionally, web page change detection methods have been developed to refresh the web document by detecting the structural, presentation and content level changes in web documents. These change detection methods help to detect whether version of a web page, existing at Search engine side has got changed at Web server end or not. If it has got changed at Web server end, the WebCrawler should replace the existing version at Search engine database side to keep its repository up-to-date.