CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

World Wide Web (WWW) is a repository of web pages which are interlinked hypertext documents. Web page consists of contents as well as links to related pages, these links are called hyperlinks. Users use Internet to access these web pages. Web browser is used to view the web pages that contain information in form of text, videos, images and multimedia and also navigate them using hyperlinks known as Uniform Resource Locators. Tim Berners-Lee proposed “using hypertext to link and access information in 1990”. Following his ideas the websites are created using hypertext markup languages and are connected through Internet.

India is the third largest country of Internet users in Asia after China and Japan. WWW has grown in size many folds, distributed all over the world on web servers and is still growing at phenomenal rate. It is difficult to search information from such huge collection of web pages and documents that are scattered over Internet. It is challenging for the users to retrieve information even after knowing URLs as Web is constantly changing and so are the URLs. Web users use search engines or other information retrieval tools to find information from WWW. Web users require Web information retrieval support systems rather than Web information retrieval systems.

The Internet is a collection of computer networks that uses Transmission control protocol and Internet Protocol to link computers all over the world. It is network of networks that consists of wide array of private, public, academic, business and government networks linked by electronic media, wireless media and optical networking technologies. The Internet has created challenges for society as well as for the technology used for the Web. Use of the Web has raised important questions in the
fields of censorship, privacy and access to information. On the technological front there is need to scale the applications to store the resulting large databases and heavy network loads. The solution to above problems lie in the disciplines of distributed systems and information retrieval. Earlier when a user requires information on the Web, he either had to know the exact location of the documents or he had to navigate patiently from one hyperlink to another hyperlink in hopes of finding his desired page. As the size of the Web grew such type of navigation became impossible. Web Crawlers have made the Web easier to use for millions of people. Web Crawlers can be considered as a Web service to assist users in Web navigation. The web graph describes the directed links between pages of the World Wide Web. The web graph is a directed graph, whose vertices correspond to the pages of the WWW, and a directed edge connects page X to page Y if there exists a hyperlink on page X, referring to page Y.

The Internet carries large amount of information and services, such as hypertext documents, the infrastructure to support email services, file sharing and telephony services. The Internet is a fault-tolerant, robust communication system using computer networks. Internet begins with the development of electronic computers in the 1950s. Concepts of packet networking started in countries like United States, Great Britain, and France. ARPANET, first network to use Internet protocol was developed in 1960. Storage plays an important role in developing of the Internet. World Wide Web is a large mineral memory. Modern mineral memory allows fast reproduction of the work at low cost with minimum human effort. The Web has transformed the world's economy and will continue to do so. The World Wide Web provides instantaneous access to vast quantities of information. With the increase in performance improvements in bandwidth and device latencies, the users continue to demand more speed while the content providers’ demands increase in bandwidth. Good response-time is essential for user satisfaction, productivity and there is significant incentive for content providers to provide effective responsive time.

There is the problem of abundance in modern working day, there are various routine tasks like dealing phone calls, internal memos, external post and e-mails [1]. Human race is living in the era of information explosion, with information being measured in magnitudes of tera bytes. The World Wide Web contains about 170 terabytes of
information [2]. During the earlier days of World Wide Web, information was found by scanning the links collected manually. This manual sorting was done based on some criterion. There were no automated web search engines. In fact web search engines were not required as there were only thousands of Web pages. In today’s era there are thousands of millions of web pages over the World Wide Web. The problem of search engines is the size of Web and rate of change of the Web. Currently there is no search engine available that is indexing more than one third of the available Web [3]. As the number of pages are increasing it is important to focus on the most “valuable” pages, as there are no search engine available that will index the entire web.

The Internet requires sets of protocols to function properly. A protocol is a well defined standard for connection, communication, and data transfer between two computers over a network. Some key protocols used over the Internet are HTTP, HTTPS and FTP. The HTTP is Hypertext Transfer Protocol, used for transferring web pages over the Internet while HTTPS is Hypertext Transfer Protocol over Secure Socket Layer, it is used on web pages that ask for personal information such as a password etc. Website using HTTPS have an SSL certificate issued by trusted third party installed on the server. When a web page is browsed using HTTPS, the details of the SSL certificate can be checked. FTP is File Transfer Protocol. It transfers files over the Internet. FTP is used by web developers to publish new version of the website. HTTP is for displaying the file in your browser while FTP is for transferring file from one computer to another computer. FTP transfers the files from one computer such as a web server to remote computer.

TCP/IP is the Transmission Control Protocol and the Internet Protocol also called the Internet model was originally called DoD model, because the development of the model was funded by an agency of the United States Department of Defense i.e., DARPA. TCP/IP specifies packeting, addressing, transmitting, routing and receiving of data at the destination. There are four layers that are used to sort all related protocols according to the scope of networking. They are the link layer, the internet layer, the transport layer and the application layer. The link layer contains communication technologies for a single network segment. The internet layer connects two or more hosts across independent networks. The transport layer handles host-to-host communication. The
application layer provides application data exchange between processes. The TCP/IP protocols are managed by the Internet Engineering Task Force.

A web browser is a software application for retrieving, presenting and traversing information on the World Wide Web. An information resource is given by a Uniform Resource Identifier or Uniform resource locator. An information resource may be a web page, video, image or any other content. Hyperlinks present in web page enable users to navigate content related to resources. The most popular major web browsers are Internet Explorer, Firefox and Google Chrome. The function of a web browser is to fetch information resources to the user, allowing them to display the information and then allow for access to other information.

The World Wide Web is a system of interlinked hypertext documents that are accessed by the Internet. Web browser views web pages as text, videos, multimedia and navigate between them via hyperlinks. Tim Berners-Lee, inventor of the Web in 1989 wrote a proposal that became the World Wide Web. The terms Internet and World Wide Web are not the same. The Internet is a global system of interconnected computer networks. The web is one of the services that run on the Internet. It is a collection of text and other resources, linked by hyperlinks accessed by web browsers.

1.2 Hypertext and the Web

Web is a kind of hypertext system that is collection of documents related by hyperlinks occurring in the body of each document. When a hypertext document is accessed the user can choose to follow the links embedded in the documents to other linked documents. This concept of hypertext is similar to concepts originally perceived by [5][6][7]. WWW is decentralized across administrative boundaries. Publishers of web documents can easily create and publish documents that anyone could read instantly. The structure of the Web encouraged publishers to include links in the documents that crossed boundaries. In the present scenario documents are spread all around the world and are created and used by millions of authors, readers and users. Due to large size of the Web there are many problems such as finding resources in such a distributed environment and the problem was more difficult to solve because the size of Web continues to grow. Solution to this problem is to combine hypertext with search.
Performing this task means changing the conventional idea of search defined in information retrieval. Information Retrieval is the broad area of computer science that deals with retrieving information about a specific subject from a well defined collection of data objects. Information Retrieval deals with satisfying the need of the user. Information Retrieval system must interpret the contents of the documents and rank the documents according to relevance [68]. Information Retrieval techniques were also published before the invention of the WWW, but these techniques are not useful for the Web but are useful for small articles such as newspaper articles or book catalogs in a library. The reason that these primitive information retrieval techniques are not useful for the Web as it is massive; it changes rapidly and is spread across various distinct geographic locations [69]. The major problem in IR is to find a relevant selected set of documents from a large collection in response to a query of searcher. The results of a query are a relevant subset of documents from a larger collection. Most of the research work in Information Retrieval focuses on systems for storing and viewing documents and methods for query processing [4]. Earlier the use of powerful retrieval systems had been restricted to a less numbers of sophisticated users. Also research in Information Retrieval focused on expert users who could use powerful complex query systems. But now a day large number of end users use sophisticated IR techniques. Search engines expose large databases to large groups of end users. The search engines provide an interface where end users pose simple queries which would otherwise be complex. Web crawling is the process used by search engines to download pages from the World Wide Web.

According to Brooks [70], there are two types of web firstly the closed web secondly the open web. The closed web consists of high-quality controlled collections of web pages on which a search engine can fully trust, while the open web consists of web pages on which traditional IR techniques and concepts are not effective. The threat that open web poses is search engine spamming. The search engine spamming is malicious attempts to get an undeserving high rank as the result of retrieval. This is new branch of Information Retrieval also called Adversarial Information Retrieval. Adversarial IR is retrieving information from collections of web pages in which a subset of web pages from the collection has been manipulated to influence the algorithms. The vector space model [71] and Term Frequency –Inverse Document Frequency similarity measure [72]
are used for identifying web documents in a collection of web pages that are relevant. However there is another technique called link Analysis. Link analysis cannot be done on traditional information repositories but can be done on the World Wide Web to exploit links and extract useful information from web pages. Web pages come without publishing costs which is the reason for the success of the web and also the demerit that searching information on the web is more difficult. A small program creates large number of web pages to maliciously increase the citation counts. Also Web involves profit seeking business; attention seeking strategies evolve in response to search engine algorithms over the Internet. So the manipulation can be done on any evaluation strategy which counts replicable features [73].

Major factors that are responsible for delay in web experiences are heterogeneous network connectivity, congestion in networks or servers. Researchers have considered the problem of improving Web response time for effective web usage. Some researchers tried to improve web response time by increasing bandwidths while others used the existing infrastructure more effectively and efficiently. Web caching helps to reduce network usage and loads on server thus improving web response time for the user. Pre-loading Web objects into local cache can reduce web response time. Pre-loading implies the transmission of data over a network that can be performed where a cache is present, including client browsers, CPU caches or Web server caches. Pre-fetching more specific form of pre-loading is also used. Pre-fetching is the (cache-initiated) retrieval of a resource into a cache in the anticipation that it can be served from cache in the future. Pre-loading is broader term encompassing both pre-fetching and pre-pushing, in which content is pushed from server to cache.

1.3 Search Engines

The web is considered as large distributed database and it does not offer any indexing and searching functionalities. The searching and indexing tasks for the web are handled from specialized web applications called search engines. There are two types of search engines, the search engines that maintain their own database and Meta search engines that perform searches by facilitating services provided from other search engines. The search engines that maintain their own database use web crawlers to crawl as many pages as possible while meta search engines answer to the search queries by combining
query results from other search engines, with no need to keep their own database with updated information. Search engines are the information retrieval tools. Search engines explore and download web documents from WWW that brings high quality documents so that the user gets the required information within time. The documents are displayed in order of their relevance. The web is dynamic in nature so search engines should update their repository. Search engines execute multiple crawlers that run in parallel, to download web documents. While working in parallel, crawlers face challenging problems such as web page change detection, overlapping, quality, network bandwidth that needs to be addressed. Search engines employ ranking algorithms to meet the objective of downloading quality web pages. Search engine keep its database up-to-date with respect to the web pages maintained at Web server end. It is important to find whether a document has changed or not.

The publicly available search engine includes the HTML interface where the users submit their queries and there is the mechanism for serving these queries. The publicly available search engines also include a search mechanism to facilitate queries in search engine tables and there exist procedures for ranking the results. The search is based on keywords. There are different methods for performing a query-oriented ranking on the web pages, so that the most relevant pages are returned first. The database stores all the crawled data from the web crawlers. The search engine queries the database in order to respond any user’s request. The URLs are stored in the database and then they are sent to downloader for further downloading. An updated database would mean better search results. A large database can give more complete answers when compared to a smaller database. The search engine databases have size of several thousands of terabytes. Search engine database must have following functionalities: fast searching for the pages, ability to find all the words that are included in a web page, ability to quickly delete, update or insert new records, ability to serve thousands of users concurrently, ability to expand in more than one hard disk drives. For this reason the search engine databases are usually distributed in several servers. Google Bigfiles are one of the few that are publicly described and are the database used for the Google search engine. The system enables the fragmentation of the database to several files that can be hosted in different machines. The fragmentation system is called Bigfiles. The size of files system is up to $17 \times 10^9$ Gigabytes. The database contains the repository, a table with the
complete text for every html page crawled. Lexicon is a table that keeps the most important keywords from all the web pages and assigns them an identification number. Document table contains specific information for every stored URL i.e. the PageRank. Hitlist implements the M:M relation between the document and lexicon table. Hitlist table also contains a rank. Forward index enables fast searching for web pages related to a keyword. Reverse index enables fast updating of the database with data from new or updated web documents.

### 1.4 Web Crawlers

All search engines maintain their own central database of web pages. In response to the user query the search engine creates indexes for the repository. Web Crawler also known as spider, robot, and web pot is a program that traverses the Web and collects web documents. Web Crawler starts from seeds URLs and download web document for the seed URLs. It extracts new links present in these downloaded documents. The extracted URLs are then checked whether they are already been downloaded or not. Once it is ensured that the documents have not been downloaded, URLs are reassigned to crawlers for further downloading. The process is repeated till no more URLs are left for downloading. Millions of web pages are downloaded per day by a crawler. Web crawler consist of scheduler, multi threaded downloader, queue for URLs and storage for storing text and metadata.

![Figure 1.1 High-level architecture of a standard Web crawler](image)
The web crawling system is responsible for maintaining the search engine database and incorporating the changes from the web. Multiple instances of this component run in parallel connected with very high bandwidth, in order to overcome the processing bottleneck. The web crawlers download and process web pages from the web, by using HTTP protocol. The web crawling system consists of two subsystems: the downloader and the processor. The downloader downloads the web pages from the Internet and sends them to the processor for further processing. This process uses HTTP protocol to download pages from the web. The downloader issues HTTP-GET requests to download web pages. The pages are then forwarded to the processor for further processing and integration in the database. The downloaders follow several good practices in order to download web pages from web servers. Some of the optimization techniques applied to the downloaders for overcoming the network bottlenecks are TCP Connection reusability, Compression, Conditional GETs, Varying refresh frequency and DNS caching.

TCP Connection reusability: To save time and network bandwidth, the downloaders keep one TCP connection open for each IP address and reuse it until they download the whole contents from the server.

Compression: To reduce network traffic, web servers compress web pages before sending them to the requested client.

Conditional GETs: Conditional GETs defined in the HTTP protocol are used by downloaders. If the page is changed after a specific date Conditional GETs downloads the page.

Varying refresh frequency: Some web pages are updated every few minutes, while some web pages do not change for a whole year. Some pages are important than others. That is why search engines crawl the pages that are important and that are expected to get updated often.

DNS caching: Downloading of URLs and DNS resolutions are reasons for both network traffic and network bottleneck. Further bottleneck in the DNS servers may also happen affecting the throughput. The downloader or the web crawling component uses DNS
cache, for improving performance. Centralized web crawling strategies fail to scale when the size of the web grows. Search engines crawl a very small percentage of the web, often ignoring important pages. For this reason parallel web crawlers are used that alleviate the network and other bottlenecks.

The processor: The processor receives the downloaded web documents, processes them, and saves them in the search engine’s database. It is responsible for extracting the information from the web documents and saving them to the database. The strategy used is to extract keywords from a web page and give them rank. The occurrence and ranking for keywords are stored in the database. The processing component performs string processing in order to extract information. Parallel web crawler is special type of web crawler which is discussed in the next section.

1.4.1 Parallel Web Crawler

The parallel web crawler runs multiple crawlers in parallel to maximize the download rate and to retrieve the whole or significant portion of the Web. The major issues with parallel crawlers are quality, network bandwidth and overlapping of web pages. When multiple crawlers run in parallel they may download same web pages multiple times as one crawler may not be aware that another has already downloaded the same page. Thus overlapping of web pages occurs. Network bandwidth is saved and utilized for further downloading if the overlapping is avoided. Thus effectiveness of web crawler is increased.

The aim of web crawler is to download important web pages. While working in parallel each web crawler may be unaware about the entire collection of the web pages downloaded collectively. The crawling decisions are based on local collection of Web leading to poor crawling decisions. Web crawlers need to communicate among themselves in order to minimize overlap and to maintain the quality of downloaded web pages. Communication will consume bandwidth as well as the crawler’s time. So, strategy is needed to minimize the communication and to maintain the quality.
1.5 Objectives of Proposed Research Work and Research Questions

The overall objectives of the proposed research work are:

- To review and critically examine the literature on Migrating Parallel Web Crawler
- To identify issues in Migrating Parallel Web Crawler.
- To develop model for domain specific and incremental web crawling
- To implement the A Neural network based change detection method in migrating parallel web crawler that will yield high quality pages and detect for changes will always down-load fresh pages.
- Solution to Network Traffic problem in Migrating Parallel Crawlers using Fuzzy Logic
- To validate the proposed model.

The research questions are:

- What are the definition, processes, and techniques in Migrating Parallel Web Crawler?
- What are the relative strengths and weaknesses of the existing Parallel Web Crawler?
- If necessary, what should be the key components of a new and improved approach for effective Migrating Parallel Web Crawler based on the available literature?
- What does the existing literature say about the state of practice in Migrating Parallel Web crawler?
- What do expert analysts in the field say about the state of practice in Migrating Parallel Web crawler?
- What do novice analysts in the field say about the state of practice in Migrating Parallel Web crawler?
- If necessary, what should be the key components of a new and improved approach for Migrating Parallel Web crawler according to the state of practice?
• Can the identified key components be combined to design a new and improved approach for domain specific and incremental crawling in Migrating Parallel Web Crawler?
• Can the identified key components be combined to construct a tool to support and enhance the new and improved approach?
• Does the approach improve the efficiency of the Migrating Parallel Web Crawler, in terms of the overall amount of information collected?

1.6 Main Contribution of the Thesis

The major focus of this thesis is:

• The architecture for Effective Migrating Parallel Web Crawling approach with domain specific and incremental crawling is proposed. Domain specific crawling will yield high quality pages while incremental crawling will keep the database fresh. The crawler is implemented in Java and supports all features of the real time III tier architecture.
• In this thesis neural network based change detection method in migrating parallel web crawler is implemented. The method for Effective Migrating Parallel Web Crawling approach detects changes in the content and structure using neural network.
• Crawlers may be generously allowed to communicate among themselves or they cannot be allowed to communicate among themselves at all, both approaches put extra burden on network traffic. Here a fuzzy logic based algorithm is proposed and it is implemented in MATLAB using fuzzy logic tool box which predict the load at particular node and route of network traffic.
• The architecture of migrating parallel web crawler is validated using finite state machine. Test cases are generated for the validation of the architecture. The approach for generating the test cases through Finite State Machine, which is very reliable, efficient method and does not support for the invalid test cases.
• Empirical validation is performed to assess the performance of the proposed approach using SPSS tool.
1.7 Structure of the Thesis

The thesis is organized chapter wise as follows:

In chapter 1 World Wide Web, Internet, Search engines and Web crawlers are discussed. The evolution of and how the web has become the source of information sought by users is discussed.

In chapter 2, hyper text system, indexing and querying web pages, web characterization and distributed systems are discussed. The web crawlers, their types and web crawling issues are also discussed. Based on the literature reviewed change detection techniques have been identified, that provided the basis for the work to be carried out.
Chapter 3 provides the outline of complete architecture.

In chapter 4 the architecture for Effective Migrating Parallel Web Crawling approach is proposed with domain specific and incremental crawling strategy that makes web crawling system more effective and efficient. Domain specific crawling will yield high quality pages. The crawling process will migrate to host or server with specific domain and start downloading pages within specific domain. Incremental crawling will keep the pages in local database fresh thus increasing the quality of downloaded pages. The web crawler as a client server system is implemented based on JAVA, Apache TOMCAT as server and MySQL as database has been designed. The implemented model supports all features of the real time III tier architecture. The Net Beans 7.4 is used as Development Platform.

In Chapter 5, a neural network based change detection method in migrating parallel web crawler is implemented. The method for Effective Migrating Parallel Web Crawling approach detects changes in the content and structure using neural network. Neural network based change detection method in migrating parallel web crawler will yield high quality fresh pages and detect for changes.

In chapter 6, discussion on the crawling process is carried out using either of the following approaches: Crawlers can be generously allowed to communicate among themselves or they cannot be allowed to communicate among themselves at all, both approaches put extra burden on network traffic. Here a fuzzy logic based algorithm is proposed and it is implemented in MATLAB using fuzzy logic tool box which predict the load at particular node and route of network traffic.

In Chapter 7, the architecture of migrating parallel web crawler is validated using finite state machine. Test cases are generated for the validation of the architecture. The approach for generating the test cases through FSM is very reliable and efficient and does not support for the invalid test cases. Valid input strings are generated as test cases. Empirical validation is performed to assess the performance of the proposed approach.