CHAPTER 1
INTRODUCTION

1.1 General
The information retrieval is a process of searching the desired information into a collection of documents. The search for a particular term(s) is/are called a query (Yates and Neto 1999). The information retrieval basically deals with the representation of documents, storage of documents and access to the information inside the documents by which a user can easily access the desired information. The process of information retrieval mentioned in this thesis is related to the online information, which is available on the World Wide Web (WWW) only.

WWW (Duhan et al. 2009) is a large distributed system of interlinked hypertext documents, which can be accessed via the Internet. A hypertext document consists of hyperlinks, a concept introduced by T. B. Lee in 1990 to links other documents and information in the form of text, audio, video, images etc. These hypertext documents are viewed by using special software called Web browser. A Web browser or simply browser is a software program used to retrieve the information from the WWW and then presents this information to the user. Initially, the information available on the Web was very limited and the end user had to find the information manually. As the time passed, the number of Internet users increased from some thousand to billions and accessing of information became increasingly difficult. The studies of (Yossef et al. 2000; Bharat et al. 1998; Bharat and Broder 1999; Gulli and Signorini 2005; Wor 2013) show the estimated size of the Web and agree on the fact that over billions of Web pages is available on the WWW. According to a survey (Wor 2013), the current population of the world was approximately 7 billion out of which about 2 billion peoples uses the Internet (Wik 2013). The survey further suggests that the Web currently includes 4.16 billion pages. It indicates that the number of documents in the Web repository is increasing much faster than its users.
Although the number of documents on the WWW has increased yet the end user’s knowledge about the Web page has remained limited. All this happens because there is no way of discovering the Web pages of user’s interest unless the user has prior knowledge regarding them. Generally, a user specifies the initial URL on address bar of the Web browser to find the desired information on the Web. After this, at each step, the end user can click on some more links available on the Web page. The end user can also access the related documents being referred to on the Web page. This process of finding the information from the WWW is called browsing or navigating the Web.

Today, more than 90% of the information of daily use is found in the electronic form. As the information on Web is increasing daily, so, the number of files is also being increased daily (Tsai et al. 2006). This creates a challenge for searching the required information on the Web. There are some automated information retrieval tools which is used for retrieving information efficiently. These tools provide user interface to search the information on the WWW. The end user enters the keyword of the desired information to be searched on this interface. A list of URLs to the sites containing the information related to a specific topic is provided by such tools. Now, end user can browse any of the links from the listed URLs. This automated searching is completely different from browsing. The automated searching tools do not require prior knowledge of specifying a set of remote computers for navigation on the Internet, instead, automated search tools use pre-defined computer programs to locate the Web pages that contain the information related to the given keywords. Precision and recall are two parameters to evaluate the performance of any information retrieval tool (Kobayashi and Takeda 2000). Precision is defined as the ratio of number of relevant documents to the number of retrieved documents, whereas recall is defined as the ratio of number of relevant documents that are retrieved to the total number of relevant documents.
The information retrieval tools can be grouped into following three broad classes i.e. Web Directories, Search Engines and Meta Search Engines.

**Web Directories:** Web directories also called search directories are used to list the Web sites according to their categories and subcategories. Web directories are just enormous collections of hyperlinks to other Web sites. The Web sites in a Web directory are listed in some order (generally in alphabetic order) and users browse through them. These directories are edited by the users of Internet and URLs are submitted by the site owners only which obsolete the use of Web directories.

**Search Engine:** A Web search engine is a software system that is designed to search for information on the World Wide Web. The searched results are presented as a list of hyperlinks. The information provided by the hyperlinks may be a mixture of text, audio, video and other types of files. Search engine maintain real time information by running an automated software called Web crawler.

**Meta Search Engine:** A meta search engine is a search tool that sends user’s request to several other search engines and then aggregates the results retrieved from those search engines into a single list and displays to the users. Meta search engines enable users to enter search criteria once and access several search engines simultaneously. Meta search engines operate on the premise that the Web is too large and it is impractical for one search engine to index it all. This may save the user from having to use multiple search engines separately. This process of fusion also improves the relevancy of search results.

**1.2 Research Motivation**

Today, continually enlarging volume of the Web, dynamicity of the Web and search services to locate the information on the Web makes the information retrieval process most popular area of research. The search services are widely used to search the information from the large amount of distributed data. The
revolution that the Web has brought to information access is not so much due to the availability of information but rather due to the increased efficiency of search tools. In fact, the search services of today make previously impractical tasks as practical. But how can users find the required information from such an unstructured, unorganized and decentralized repository of information. The modern techniques of searching the information employ various techniques to optimize their search process. The major problem for searching the information is the inability of finding the exact information that the user is looking for. The surveys of various publicly available literatures (Nath and Bal 2011; Lawrence et al. 1999) suggest the following reasons as to why finding the information using the Web search services is not so successful.

(i) **Large Size of the Web:** The first difficulty is related to large amount of information available on the Web. The quality of information search is greatly influenced by the completeness and freshness of the information available in the indexed repository of the search services. But, no search engine is able to index the entire Web. A study in (Nath and Bal 2011) showed that no search engine was able to cover more than 40% of the entire Web. However, another issue of decreasing the effectiveness of a search service tool was the dynamic nature of the Web. It may result in “dead links” and “out of date” pages that were changed since the last crawl and indexing. But, even after considering such factors, the current Web service tools are not able to search the relevant information.

(ii) **Ambiguity in Determining the Relevancy:** One of the major problems in searching the information was in determining the relevancy of the query with the Web page. The page rank algorithms were based on link analysis used in determining the relevancy of the Web page. The experience showed that, in many situations they were not enough to provide the desired results. As today, the importance of time has increased, so, each user hopes to see the search results in first few links only. Therefore, the problem of ranking the search results became a big problem for the current research. Another solution for
ranking the search results was to rank them on the basis of similarity between the query terms and the retrieved Web page. But such types of solution required proper formulation of the query by the searcher and the training process, which could depict how to search and how to formulate a query.

(iii) **Dynamic Nature of the Web:** Both implicit and explicit structure of Web data was updated regularly. So the aim of searcher was to find the latest information. The articles like news services were frequently updated and could not be tracked and ranked by using link analysis techniques. Especially, due to different applications of Web based data management systems, a variety of presentations of Web documents were generated and deleted continuously. Dangling links and relocation problems arrive when domain or file names changed or disappeared. Moreover, when a new article was posted, it had no hyperlink pointing to it. Therefore, a search service must be able to track such a dynamic nature of the Web.

(iv) **Low Precision:** Low precision was caused by a lot of irrelevant pages returned by the search services. Search services returned a huge list of links in response to the user query. This list could be reduced by properly formulating the search query. However, searchers often formulated very short query that might result in large number of documents. But most of the queries were limited to single word only. Therefore, it became very difficult for any search service to identify the documents of user’s interest. Hence, lots of retrieved documents were not of user’s interest. Such type of situation was termed as low precision search. Low precision search might result in loss of user’s interest and might become a problem in finding the desired information. Such a situation was called Information Overkill. Hence, there should be a mechanism that could help the user to deal with large document sets.

(v) **Fuzzy Formulation of Queries:** The results of any search service depended upon the type of query given by the user. Very short queries for which few resources existed on the Web, created a challenge called “needle in the
haystack”. But, broad topic queries might result in a large quantity of returned results. Furthermore, the significance of a document based on the user query was a time varying concept. Sometimes, the same set of keywords might fulfill different user needs that may also vary eventually according to the context of query formulation. All this happened because of the phenomena called “Synonym” (several words corresponding to a single concept) and “polysemy” (one word with multiple meanings). Therefore, a search service should be able to handle such types of problems also.

(vi) Different Media: Media specific search services that index audio, video and images were also gaining more popularity. However, these services were not able to know about what type of media were available on a Web site. All these search services firstly crawl the HTML pages and then identify their media by looking at their extension (e.g. .GIF, .JPG or .mp3). Also, the media at a Web site that were not hyperlinked with some HTML page were not found by the media specific crawler. These problems can be alleviated if crawlers and search tools use a more appropriate protocol that helps to identify ‘what should be copied’. For example, let us assume that the Web server maintains a list of URLs and their last modified dates for all available Web pages. Then, the Crawler should download the URL, which has been modified since the last crawl and request only the pages that have been changed. This may significantly reduce the number of requests the crawler makes to the Web server and thus the Web server may save its valuable resources being wasted by servicing these requests.

1.3 Problem Statement

Today, the main challenge in front of any search service is the retrieval of latest, relevant and quality documents that can satisfy the current user needs. However a micro level critical look based on the size and dynamicity of WWW, shows that the information retrieval tools needs to address many important challenges including the following:
(i) **Document Overlapping:** Multiple Web crawlers can run and download the Web pages from different machines. Due to the huge size of the Web, multiple Web crawlers, running without communicating with each other, can download the same document multiple times. Many organizations reflect their documents on many servers to avoid the chance of server corruption (Lawrence et al. 1999; Asadi and Jamali 2004). Therefore, these parallel crawlers should be synchronized properly so that different crawlers do not visit or download the same Web page multiple times.

(ii) **Pages a Crawler should download:** No Web crawler can download all pages available on the WWW because any search engine can cover only 40% of the entire Web (Nath and Bal 2011). So, it becomes important for a Web crawler to carefully select the pages to be visited and download the important pages promptly.

(iii) **Network Bandwidth:** When the crawler collects pages from the WWW, it consumes resources belonging to other Web sites (Tan and Kumar 2002; Nath and Chopra 2013). For example, when the crawler downloads any Web page (p) from any Web site (S), the site needs to retrieve page (p) from its file system, consuming disk and CPU resources. Now, the Web page needs to be transferred through the network, which is another resource shared by multiple organizations. Therefore, the crawler should minimize its impact on these resources.

(iv) **Refreshment of crawled pages:** After downloading the Web pages into the repository of the search tool, a crawler has to start revisiting the downloaded pages again and again in order to detect the changes in the downloaded Web pages. Because the Web pages are changing at very different rates (Cho and Molina 2000; Wills and Mikhailov 1999), the crawler needs to carefully decide which page it has to revisit and which pages to skip in order to achieve high “freshness” of pages. For example, if certain page rarely changes, the crawler may require to revisit these pages less often, in order to visit more frequently changing ones.
(v) **Security:** The migration of crawler on remote network/computer and executing the code at remote site, sometime, causes severe security problems because a mobile crawler might contain some harmful code (Nath and Bal 2011).

(vi) **Post Relevancy Determination:** The relevance of any Web page can be determined after downloading it on the search tool end only. So once a page is downloaded then what is the benefit of filtering it, because all the network resources are already used while downloading those pages? (Nath and Bal 2011).

(vii) **Memory Scarcity:** If the mobile Web crawler stays in the memory then it consumes significant segment of memory, so what happens when more mobile crawlers from different search engines stay there? This can produce scarcity of memory on the remote system site because remote system has to perform its task also (Nath and Bal 2011).

(viii) **Availability of Resources:** The crawler may depend on the resources of the remote host for its processing power, but it is not possible to predict the amount of memory available to the crawler (Nath and Kumar 2012).

(ix) **Dynamicty of Web Pages:** If a Web page changes rapidly and every time a mobile crawler catches these changes and send these updated pages to search engine, then it again consumes more resources of the network including the processing of the system.

### 1.4 Objectives of the Thesis

The main focus of this thesis is to investigate the issues that are concerned with the Parallel Domain Focused crawler, the presentation of results, the search engine evaluation and the automatic text summarizer. To increase the in-depth knowledge of the subject following objectives has been set for this thesis:

(i) To develop a method for Web page change detection to identify whether the two versions of a Web document are same or not. This can help in deciding whether the document is to be downloaded or not.
(ii) To develop an architecture for Parallel Domain Focused Crawler that helps in reduction of load on the network, time consumption and improves the network utilization.

(iii) To develop a Meta Search Engine architecture that can provide relevant results and reduces the redundancy in the results.

(iv) To develop an Automatic Text Summarizer that can provide the summary to selected single as well as multiple documents by reducing the redundant results in order to improve the network utilization.

(v) To develop an Automatic Search Engine Evaluation framework that can determine the specialization of a search engine.

(vi) To implement all the models presented above to validate them.

1.5 Organization of the Thesis

The thesis has been organized into eight chapters. A brief description of each chapter is presented in the following paragraphs.

Chapter 1: It presents the general information about the search systems. It also discusses research motivation and problem statement for this thesis. Further, it sets the objectives to be achieved through the study.

Chapter 2: This chapter presents the literature survey, which was carried out for this research work. This chapter is divided into four main sections i.e. Web crawler architectures, meta search engine architectures, automatic text summarization and search engine evaluation. Each of these sections describes the publications related to them.

Chapter 3: This chapter analyses the literature presented in the previous chapter and discusses the issues and challenges in detail, which are involved in the design of Web crawlers, meta search engine, automatic text summarization and evaluation of search engine. This chapter briefly describes the architecture of each of these topics along with their merits and demerits.
Chapter 4: This chapter discusses the problem with the existing traditional and mobile crawling techniques and proposes a Parallel Domain Focused Crawler to overcome these problems to crawl the Web. It discusses the complete working and algorithm of the proposed architecture. It further evaluates the proposed crawler by comparing it with the existing crawler PMCS.

Chapter 5: This chapter discusses the problems with the traditional and existing meta search engine and proposes a meta search engine to overcome these problems. It discusses the complete working of proposed architecture along with related algorithm. The proposed meta search engine focused on the relevancy of returned results and minimization of redundant document. It further presents a comparison between the proposed and the existing MSE’s.

Chapter 6: This chapter discusses the problems available with the traditional and existing automatic text summarizers. It proposes and discusses the complete working of an automatic text summarization approach that does not use any preprocessed data. It also presents a comparison between the proposed and the existing summarizers. Furthermore, it also describes the application area of the proposed summarizer.

Chapter 7: This chapter discusses the problems related with the existing techniques used for evaluation of search engine. A search engine evaluation framework based on TLS is proposed and implemented. The proposed framework evaluate the search engine on three parameters i.e. relevance of returned results, precision of results and user efforts.

Chapter 8: This chapter presents the conclusion of the thesis and provides the scope of future work.

References: It presents the exhaustive list of references used in this thesis.

Appendix A: Provides a list of URLs used for Web page change detection.