CHAPTER-1
INTRODUCTION

1.1 WWW AND INFORMATION RETRIEVAL TOOLS

The World Wide Web had been in existence since 1993 as a result of effort by Tim Berners Lee at CERN, the European Particle Physics Laboratory [1-3]. He tried to develop a system that researchers, scientists and academician can use to exchange the information. CERN maintained information on various computers. People came to CERN to use different computers and different software, as it was a challenge to share the information between different computers. Lee [4] used hypertext markup language (HTML) in connection with Internet. The Internet is a network of interconnected computers spread over WWW. Lee developed WorldWideWeb as first web browser and editor that was capable of interpreting the HTML documents. Lee’s system allowed people to share the information independent to type of computers, Operating systems and Web browser they use.

With the time, WWW became a huge repository of hyperlinked web documents. From 1993 to 1996 the size of Web increased from 130 to 600,000 web pages which as of now is approximately and is still growing exponentially. A web document is consists of contents related to some topics and number of hyperlinks to other related web documents. The information from World Wide Web, in form of these web documents can be access using client-server architecture of the web. Users get the desired information by following hyperlinks present in web documents. Since web documents are not organized as books in a library nor are catalogued at one place; it is difficult for a user to get the desired information by exploring the hyperlinks. Moreover, the size of the WWW is growing exponentially, making it more difficult to get the desired contents by exploring the links manually. Thus, arises the need of some automatic tools then can search the information from WWW using Internet. These tools are known as information retrieval tools.

All available information retrieval tools can be divided into following three categories:

- Web Directories
- Meta Search Engine
- Search Engine
1.1.1 WEB DIRECTORIES

The web directories are also known as web portals or taxonomies. It organizes the web documents in a tree like structure where each node represents a topic or subtopic. In fact under this scheme general topics are further divided into subtopics. It helps the non-expert user to get the desired information easily through tree traversal. Yahoo, Look Smart and Open Directory Project (ODP) are some example of web directories. The Open Directory Project claims to be “the largest, most comprehensive human edited directory” [5]. As the user walk down to the taxonomy tree, the nodes represent more specific topics. Thus users can get the general information by traversing up in the tree. Following drawbacks of the taxonomy tree are observed.

- It requires human intervention to assign documents in various categories.
- Different hierarchies for the same number of topics may be there by different human experts.
- Users sometimes follow the longer path to get the relevant information.

1.1.2 META SEARCH ENGINE

Meta search engines are the information retrieval tools that do not maintain their own database repository. They pass the user query to different search engines and compile the returned results. Finally they serve to the users the compiled information in some relevant ranking order. As each search engines has its own unique index of web documents, because of its unique crawl and indexing strategy, Meta search engine ensures that no matched information related to the query are missed. Metacrawler, Dogpile, Copernic and SurfWax etc. are some of the examples of Meta search engines. The need for Meta search engine has decreased now a day.

1.1.3 SEARCH ENGINE

Search engines (SE) are information retrieval tools that maintain their own repository of web documents and index them. The repository is built with the help of special programs called crawler that downloads the documents from WWW. Crawlers extract the link information from downloaded web documents and keeps on downloading the linked web pages. (In this thesis the web pages and web documents are used interchangeably and denote the same
object). User can access the desired information by passing the query keywords to SE through user interface. The query processor module of SE retrieves the matched information from indexed repository and presents them to user in some relevant ranked order.

The earlier search engines were JumpStation, World Wide Web Worm and Repository-Based Software Engineering (RBSE). RBSE was the first to introduce ranking system to help searchers to find relevant information in 1993. In 1994, Lycos was first to use links of web pages to judge their relevancy. Alta Vista has developed the system to index the Web using thousands of crawlers working in parallel. In 2004, Microsoft changed its search technology by using the services of Looksmart and Alta Vista to its own crawler called msnbot. In 1998 Google was launched by Sergery Brin and L. Page students at Stanford University. They have proposed the architecture for a large scale hypertextual search engine \[29\]. The size of the web indexed by Google is the largest as compared to other search engines such as Yahoo and Bing.

### 1.2 MOTIVATION

The commercial search engines index large portion of the Web. When users pass their query to such a general search engine, it searches its indexed repository and presents a list of large number of web pages in response to the query. However, this list of web pages may contain irrelevant web pages also. Generally, the user is not interested in such a large set of results, as it is not possible to go through all the web pages returned; leading to the problem of information overkill \[6,7\]. This happen because the web search engine try to cover the whole entire Web and response the user query with all possible topics \[8\].

Moreover, it has been observed from the users search trend \[9\] that they generally do not traverse beyond the third screen of the displayed results. Thus, it is a challenge for a search engine to provide the more relevant document at higher positions. The solution of this comes out in the form of focused search engine.

#### 1.2.1 FOCUSED SEARCH ENGINE

Focused search engine aims to concentrate on the quality of the information rather than the quantity. It searches and retrieves only the subset of the Web, relevant to specific topic of interest. The focused search engine using various techniques tries to judge the relevancy of
matched web pages in respect to the user query. The challenge for focused search engine is to devise a mechanism to decide the relevance of a web page w.r.t. a specific query.

1.2.2 ROLE OF CONTEXT IN FOCUSED SEARCH ENGINE

Even though the focused search engine judges the relevance of web pages in respect to the query, still provide some irrelevant results in the result list. This problem may be due to the reason that user query context may not be clear from the query keywords provided. Thus, query keywords can be ambiguous in their meaning, leads to irrelevant results. For example, the query keyword ‘Bank’ has various meanings or senses such as river of a bank, commercial finance bank, a sloppy road etc. depending upon their usage.

Thus, there is a need for a focused search engine that can search the repository in respect to specific context of the user query and can display them in order of relevance such that highly relevant web pages are displayed at top positions.

Thus main focus of the work presented in the thesis is to devise a mechanism that can provide more relevant documents w.r.t user query. We have designed an approach for a focused search engine that not only considers user query context but also is capable to give high precision results in top ‘k’ position.

1.3 MAIN CONTRIBUTION OF THE THESIS

The focus of the thesis is to investigate the issues in focused search that deals with the context of the user query keywords and tries to judge the relevancy of web pages in respect to various contexts of words present in them. Following contributions were made to improve the working of focused search engine.

- A novel mechanism to judge the relevance of the web pages in respect to various contextual senses of words has been developed.
- Back-links are considered as important source of information and a method for back-link extraction and relevance evaluation has been developed
- Improvement in the index structure is done so as to store the contextual information of words along with other useful information at same place.
- A novel architecture for context based focused search engine is designed that helps to display more relevant web pages in top ‘k’ results in response to a user query.
1.4 ORGANIZATION OF THE THESIS

The thesis is organized in eight chapters. A brief overview of each chapter is as follows:

**Chapter 1:** This chapter covers the introduction about World Wide Web (WWW), Internet and search engine and web crawlers. The evolution of WWW and various search tools has been discussed.

**Chapter 2:** This chapter focuses on basic concepts of various information retrieval tools. It describes the architecture of a general search engine. It provides a detailed review of existing techniques in the field of focused search, the link structure of the Web and the various existing ranking algorithms. A study on user’s search trend is also provided. In this chapter finally based on the literature review, major challenges and limitations in existing approaches are identified.

**Chapter 3:** In this chapter, a novel architecture for context based search engine has been proposed. The proposed architecture uses web pages as well as their back links URLs. The architecture consists of three main layers. Bottom layer collects the downloaded web documents and back links in local repository and maintains the index of local repository. The middle layer computes the relevance score of each matched result searched by query processor and then pass the ranked list to upper layer, to display to the user. The detailed working and algorithms of bottom layer components has been provided.

**Chapter 4:** In this chapter the proposed mechanism to judge the relevance of web pages based on the contextual senses of words has been discussed. Further, in the chapter, it has been shown that how the contextual characteristics can be used to evaluate the relevance of the web documents. The performance evaluation of proposed algorithm is also discussed.

**Chapter 5:** In this chapter a proposed ranking mechanism based on the contextual senses has been elaborated. The performance evaluation of proposed algorithm is discussed using standard metrics precision and recall.
**Chapter 6:** This chapter discusses the effect of back-links on ranking results. It also elaborates mechanism for back link extraction and its relevance evaluation in detail. Algorithm design, its implementation and results are also discussed.

**Chapter 7:** This chapter presents design, implementation and testing results of prototype built for a context based focused search. Implementation of the user interface is discussed in detail under this chapter.

**Chapter 8:** This chapter summarizes the work done in the thesis. It highlights the key contributions of the research work. The scope for future work in this research is also suggested in this chapter.