CHAPTER 1

1. INTRODUCTION

1.1 INTRODUCTION

This research is concerned with the study and analysis of search engines, storage mechanism of the search engine repository and the algorithms used for search and it proposes a traffic adaptive optimum updating scheme for updating the search engine repository. It also proposes a semantic search process to improve precision in search results. Information retrieval is that branch of computer science which studies the retrieval of information from a collection of written documents. The retrieved documents aim at satisfying a user's need expressed in natural language. Information retrieval applied to huge volumes of data in the web servers distributed worldwide can be termed as web information retrieval (WIR). Web information retrieval is highly popular and presents a technical challenge due to its heterogeneity, bulk size, dynamic nature, duplication, high linkage, and wide variance in users. Many methods exist to retrieve information from the web, with each method having its own purpose of existence and mode of operation such as web directories, search engines etc.

A Graphic, Visualization and Usability Center (GVU) study showed that about 85% of people use search engines to locate information (GVU, 1998) and many search engines consistently rank among the top sites accessed on
the web (Media Metrics, 1999). With users not able to retrieve relevant results to the query, search engine research has a long way to progress before users find what they originally require. The research focus of the thesis is to maintain the search engine repository up-to-date with optimum resource utilization and improve the relevance of the search results. This chapter introduces web information retrieval and the relevant areas such as search engine, web crawlers, coherence and relevance problems.

1.2 WEB INFORMATION RETRIEVAL

Web information retrieval deals with a technique for finding relevant web pages for any given query from a collection of documents. The quality of pages varies widely and thus being relevant is not enough in web information retrieval. According to Huang (2000) the goal of WIR is to return both high relevance and high quality pages. The big challenge to attain the goal of WIR is the heterogeneous character of the web and ill-formed queries. Efficient indexing of web pages and extended search process by search engines is required to imbibe the user’s interests as also to accomplish the goal of WIR.

The enormous number of existing web pages, their rapid increase and frequent updating makes straightforward indexing, whether it be by human or computer-assisted means, a seemingly impossible, Sisyphean task (Kobayashi and Takeda, 2000). According to Arasu, Cho et al (2000) the web requires new techniques, or extensions to old ones, to deal with gathering information, making index structures scalable and efficiently
updateable. The fundamental architecture of various search engines being the same, each search engine differs due to the efficiency of individual components resulting in different quality of results. Among other components, web crawler component of search engine plays a vital role in building and maintaining results of search engines. Section 1.3 and section 1.4 discuss about the components and process responsible for lack of coherence in search engine repository and lack of relevance in the search results.

1.3 WEB CRAWLERS AND COHERANCE PROBLEM

Crawlers are small programs that browse the web on search engine’s behalf to build and maintain its repository. The programs are given a starting set of URLs whose pages are retrieved from the Web. The crawlers extract URLs appearing in the retrieved pages and use this information to determine what links to visit next. They continue to visit the web until local resources such as storage and bandwidth allotted are exhausted. The crawlers revisit the retrieved web pages to check if the contents of the pages are modified in order to update the repository.

As the web is dynamic in nature, the size of the web increases exponentially thus increasing the percentage of changing documents. Glover (2001) defines the problem of keeping the repository contents consistent with the pages to which they refer as coherence problem. Limited bandwidth, no effective method to detect the varying frequency of change and frequent updating of the web pages at the rate of 40% every month (1999) are the limiting factors for web crawlers to build and maintain the search engine
repository up-to-date. This leads to lack of coherence where the content of the
search engine repository is different from the database present in the web
space.

Lack of coherence is also created due to the movement of the web
page away from the previous location or removing the page from its location,
which is termed as deadlink. Lam (2001) states that there is a high chance of
dangling links in the web. According to the bi-annual WWW surveys
conducted at the Graphics, Visualization and Usability center of Georgia
Institute of Technology, broken links or dead links are registered as the
second most commonly experienced problem with the web. Dead links are
annoying for users and are one factor used to describe the effectiveness of a
search engine- the fewer the dead links the better (Glover, 2001).

Study and analysis of the components of search engine, which are
responsible for maintaining an up-to-date repository, has been carried out. An
autonomic computing based updating scheme is proposed to keep the search
engine repository up to date with the global database.

1.4 SEARCH PROCESS AND RELEVANCE PROBLEM
Search process of search engines involves online data retrieval component
and offline data acquisition component. Web data acquisition components
gather data from the web pages collected by the web crawlers and build the
index. The data present in the index is used to retrieve the search results for
a given query. Therefore the index should maintain required information
relating to web pages, which will enhance the search process to provide relevant search results.

But Finkelstein et al (2001) concludes that the striving of web search today to provide users with relevant search results is one of the most challenging problems of search engines. When ambiguous or general query terms are used, generic search engines return results that lack relevance to the users' query. When more specific query is given to search engines, the relevance of the results is improved. In both the cases the internet search engines provide a first level of information filtering and the users perform most of the relevance filtering based on their interest.

At present the results for a given query are identical, independent of the users or the context in which the user made the request. The interest of the user determines the relevance of search results and it differs from person to person. Search performed by information professionals differ from the one performed by general public. Since advanced searchers make up only a small portion of search engine traffic, the type of the user and the kind of keyword continue to contribute the relevance of the search results.

The second goal of the research is to improve the relevance of search results. This proposed by mootng the idea of additional entity headers to the existing HTTP headers to carry the contextual information of the web pages. The contextual information is then used to provide more relevant search results to users.
1.5 OBJECTIVES

Crucial for the web information retrieval is the availability of a suitable search engine. The efficiency of the search process depends on the way the information about the web pages (metadata) are structured and stored, and the process using, which they are retrieved. Towards this, search engine repository and search algorithms are developed. The main objectives of this research have been to:

1. study and analysis of search engines and related techniques such as storage mechanism of the search engine repository and the algorithms used for search.

2. design and develop a representation of the metadata of web pages in the search engine repository.

3. design and develop a representation of the link between the updation of web pages and meta data updation in the search engine repository.

4. design and develop an algorithm to improve the relevance of the outcome of the search engines.

5. design and develop a software prototype to prove the concepts of the above representation and algorithms.

1.6 THESIS ORGANISATION

The thesis analyses the components and factors that are responsible for maintaining an up-to-date search engine repository. It also analyses the drawbacks in the search process to deliver relevant search results to the user query. The proposed architectures for autonomic updating scheme and semantic search process are explained. Finally, the results of the proposed
architectures are tested on a web workload model for improved freshness of search engine repository, more relevant search results and reduced load on web servers and network bandwidth. The results are compared with the performance of previous methods and processes. Figure 1.1 depicts the main topics covered in the thesis.

Figure 1.1 Topics covered in thesis

Chapter 2 addresses and examines the approaches that are used to update the search engine repository, and compares the existing algorithms used in the updating process. It also brings out the necessity of an alternate approach to reduce the needless requests of the crawlers in updating the search engine repository.
Chapter 3 reviews in detail the existing search process and its online and offline operations. It analyses the reasons behind less relevant search results produced by the search engines. The chapter concludes with the need for context in the search process of the search engines to enhance the relevance of the results.

Chapter 4 proposes an updating scheme based on autonomic computing. This chapter discusses the autonomic characteristics of the scheme, delta encoding and top down approach used in the design. The data structures designed to implement the concept is also explained in the chapter.

Chapter 5 explains the proposed concept using state diagrams for AC node and server operations. Algorithms designed to implement the concept are given. It discusses the testbed used to validate the concept and results generated. The results are analysed for effective implementation of the concept.

Chapter 6 proposes a semantic search process to improve the relevance of search results. The chapter discusses the ways in which context has been introduced in various phases of the search process, the design issues and the algorithms used. It depicts the results of the semantic search process for different keyword, context and category combinations. It also analyses the additional memory consumed by the semantic index.
Chapter 7 describes the evaluation on the performance of the proposed concepts. It compares the performance with the existing techniques and analyses the improvements achieved by the proposed concepts.

Chapter 8 summarises the contributions made in the research. It discusses the limitations in implementing the proposed work and concludes by bringing out the benefits of implementing the concept. It also proposes future directions in which this research work can be further extended.