CHAPTER – 1

INTRODUCTION

1.1 Introduction

In every field of human activity information is the key component. Whether it is research and development, business and industry, for general people, the information has to be acquired, processed, stored, retrieved and disseminated for information. Information generation, dissemination, transfer and communication take place between people through channels and in a variety of contexts and environments. In an information-driven society and economy, information about information reigns supreme.

The Internet is completely changing and revolutionizing the way that people access information, and has opened up new possibilities in areas such as digital libraries, general and scientific information dissemination and retrieve, education, commerce, entertainment, government and health care. It is also well known that the Internet is growing rapidly, perhaps at an unprecedented rate. The growth of the Internet accelerated with the invention of the World Wide Web (WWW) Hypertext Transfer Protocol (HTTP) and the creation of a number of Web sites on the Web. With the creation of a large number of Web sites, the freely available information on the Web is increasing rapidly. From its origins in 1991 as an organization-wide collaborative environment at European Council for Nuclear Research (CERN) for sharing
research documents in nuclear physics, the Web has grown to encompass diverse information resources. The Web is a gigantic library, as it includes more than 19 billion pages as announced by Yahoo in 2005 (Pandia Search Engines News, 2007). Based on Netcraft’s data it is estimates some 29.7 billion pages on the Web in February 2007. The ISC Internet Domain Survey has identified 433,193,199 numbers of hosts in January 2007 and again 600,000,000 numbers of hosts in July 2007 (Internet Systems Consortium, 2007). All this leads us to estimate that the number of Web pages must be somewhere between 15 and 30 billion. It is also worth to mention here that the numbers of Internet users are also increasing with the increasing of Web resources. According to a report announced by the Internet World Stats, the total number of Internet users as on December 2007 is 1,319,872,109 (Miniwatts Marketing Group, 2007). Add to this, the fact that the Web lacks the bibliographic control standards we take for granted in the print media. Instead of a central catalogue, the Web offers the choice of dozens of different search tools, each with its own database, command language, search capabilities and methods of displaying results. This leads to the development of different search engines and subject directories. The prime approach to search the Web is the search engines. A variety of systems have been developed to provide effective access to these resources. As Internet become more commonplace, the need for implementing the ability to search for content has become more important.

Additionally, Web search engines continue to attract large numbers of Web searchers. The Web search engines are the most popular Web sites in terms of number of visitors to a particular Web site. This implies that most users view search engines as the best method available for finding information on the Web. According
to ComScore Media Metrix report, some 84.1% of Internet users use a Web search engine every month to find information (ComScore Media Metrix, 2006). It is established that search engine use is the second most popular Internet activity next to e-mail. The Internet and the WWW represent significant advancements for the retrieval and dissemination of scientific and other literature and for the advancement of education. The only problem is to find out the relevant document which one is being looked for. And to cope up with this problem, the search engines are introduced in the cyber world.

Though many Internet-enabled applications and services are available today, the primary use of the Internet (other than e-mail) is for information retrieval. With the advances in easy to use Web page development tools, individuals have joined organizations in publishing information on almost any topic imaginable which is accessible to millions of people easily. Of course, with such a diversity of content, and with the enormous volume of information on the Internet, retrieving relevant information is a difficult task. However, there are four different methods for locating information on the Web. First, one may go directly to a Web page simply by knowing its location. Second, the hypertext links emanating from a Web page provide built-in associations to other pages that its author considers to provide related information. Third, narrowcast services can push pages at users that meet a particular user profile. Fourth, search engines allow users to state the kind of information they hope to find and then furnish information that hopefully relates to that description.
Millions of people use Web search engines everyday to find information. Therefore, the performance capabilities and limitations of Web search engines are important and significant areas of investigation. A critical research area is the need for a greater understanding of the differences in Web search engines for the same queries. A true and fair evaluation of a search engine's performance and database is very much dependent on the evaluator being familiar with the subject matter and the search facilities offered by the search engine involved. Search engines do not index all the documents on the Web, nor do they all index the same sites or in the same fashion. In what ways do Web search engines differ from each other? Currently, we know that Web search engines differ from one another in three primary ways—crawling reach, frequency of updates, and relevancy analysis. Measuring the information retrieval effectiveness of World Wide Web search engines is costly because of human relevance judgments involved. However, both for business enterprises and people it is important to know the most effective Web search engines, since such search engines help their users find higher number of relevant Web pages with less effort. Furthermore, this information can be used for several practical purposes.

With the application of Information Communication Technology (ICT) in the library services, today’s libraries are providing various services to their patrons. As the present day libraries are heading towards the digital library approach, the need is arises for the library professionals to be well aware with the various available web information services. A present day library patron may approach to a library or an information centre to meets his/her various information needs. Today’s library professionals have to depend more or less on different Web search services to meet
the information needs of its diverse users as well as for the library itself for various in house purposes. Therefore it is very much important for the library professionals to know the various Web search services and their search strategies available in the cyber world.

1.2 Objectives of the Study

Over the past years, the Web search industry has undergone profound changes. Heavy investment in research and development by the leading Web search engines has greatly improved the quality of results available to searchers. Earlier this year marked the fourth major entry into the search market with the launch of MSN’s search index. The rapid growth of the Internet, coupled with the desire of the leading engines to differentiate themselves from one another gives each engine a unique view of the Web causing the results returned by each engine for the same query to differ substantially.

Keeping in minds the importance of Search Engines as the information retrieving tools, the study is being carried out with the following objectives:

- To study the basic components, utility of search engines as incorporated in retrieving Internet resources;
- To study the history and development of search engines over time and the most used search engines on the web;
- To alert the Internet users to the problems that arises because of inconsistent behaviour of the search engines over time;
• To evaluate the search performance of most used search engines in terms of precision and relevancy of search results;

• To suggest appropriate search strategy to extract pinpointed and exhaustive search results.

1.3 Scope of the Study

The performance and capability of web search engines is an important and significant area of research. Millions of people worldwide use web search engines every day. There are millions of pages and billions of words on the web. The search engines and directories help to sift through all those billions of 1's and 0's to find the specific information one needs. During the study, it has been realized that search engines results are not stable. The web keyword search engines don't provide enough guidance to adequately focus the search to narrow down the results to manageable relevant subsets.

The motivation of this study is based on the fact that identifying the most effective Web search engines satisfying the current information-needs is important both at a personal and a business level. Among others, this information can be used for (a) finding more relevant documents with less effort; (b) motivating search engine providers for higher standards; (c) implementing custom-made meta-search engines.

Given the above, the need is clear to study and familiarize with a variety of search tools and to develop effective search techniques, to take advantage of the resources offered by the web without spending many fruitless hours in a sea of irrelevant
information. As I'm working in a library having electronic environment, I have had experiences with different search engines and their searching strategies. In my working experiences with the web search and interaction with different users, I have faced different problems of some of the most used search engines, which needs a thorough study to understand its strategy. The performance and capabilities of web search engines is an important and significant area of research. Millions of people worldwide use web search engines every day. The goal of the present study is to conduct a large-scale study to measure the performance and efficiency of search engines on the first result page across the five most popular search engines, at specific points in time using a large number of selected queries. The web search engines included in the study are AOL, Ask, Google, Live, and Yahoo.

1.4 Hypothesis

The study assumes that search engines are the most important Internet activity on the web after the e-mails. There are numbers of information and literature about search engines and search techniques, but we will find very little on their search performance. On the above assumptions, the following descriptive hypotheses are formulated for verification under the study:

i. There is a significant difference in the number of relevant and irrelevant links retrieved by different search engines.

ii. No two-search engines are identical in terms of their coverage and search results.
iii. There is significance fluctuations in retrieved results among top rated search engines.

iv. There are significant differences among the top rated search engines in their retrieving performance.

1.5 Methodology

The study includes searching of different primary, secondary, tertiary including micro and macro literatures. In order to have a comparative study of the selected search engines, some standard evaluation criteria have been incorporated in this study.

For the study, the data have been collected by searching different Internet resources. Various studies have also conducted relating with the research by doing different surveys. Furthermore, discussions with experts, specialists, Internet surfers and library users have also been conducted.

The set of selected query terms are used in the five selected web search engines to evaluate the search engines in terms of their coverage, relevancy, stability, time factor, and uniqueness of results.

For different surveys and for collecting data, one questionnaire was designed and distributed among the internet users of different groups within IIT Guwahati. For the study, one thousand keywords were collected from the articles published in both national and international journals in the field of library and information science in
the year 2006 (Annexure IV). After deleting the duplicate and common keywords from the list of one thousand collected keywords, fifty keywords have been chosen for this research works, based on their weightage and relevancy.

The collected data are evaluated and tested by using the SPSS 12.0.1 software to derive mean, standard deviation and to measure the significance of difference among the variables.

1.6 Limitations of the Study

With the development of science of technology, new branches of knowledge are also emerging which is influencing in the designing of new web sites in world wide. As the volume of indexable web pages is increasing day by day, the coverage of search engines is also increasing. With the increasing demand of search engine among Internet users, the search engines are also providing various facilities along with the search. In the present day search world there are thousand of search engines available on the web. Among these search engines very few have been listed as popular by some search engine rating agency. Based on their reports, five top most search engines have been selected for this research works. For the convenience of self evaluation of search results, the query terms were selected from the areas of library and information science. Although the efficiency of a search engine is measured mainly by its recall and precision value, in this study only the precision value has been evaluated, as it is difficult to assessed the number of relevant pages for a particular quarry in a search engine indexes. To evaluate the recall value, it is essential to know the number of relevant pages for a query in a search engines.
indexes. According to various search engines rating agency, the surfers used to look at only the first ten hits for a particular search. Based on this assumption, the first ten results have been used for evaluation for this research works.

1.7 Statement of the Problems

Searching desired information on the Internet is one of the most common ways the Internet is used. A user of the Internet or the store at times seeks certain information, which he may not know to solve a problem. He therefore, has to express his information need as a request for information in one form or another using any search engine. Search engine then tries to infer and retrieve the relevant documents by performing the matching of query to the surrogates of documents and present the likely relevant documents to users in the form of hits list. Each search engine has a corresponding database that defines the set of documents that can be searched by the search engine. Usually, an index for all documents in the database is created and stored in the search engine. Today search engines have become one of the most critical applications on the web, driving many important online businesses that connect people to information. As the web continues to grow its size with a variety of new data and penetrate into every aspect of people’s life, the need for developing a more intelligent search engine is increasing. This needs leads to the study of the searching strategies and the evaluation of the existing most used search engines on the web.

The key for finding relevant information on the Internet is related to the ability to use various search engines. On searching the Web, people usually rely on only a few
search services such as Google, Yahoo, Live or Ask. At present, Google is the dominant service among them. For information retrieval purposes most people use either Google directly or one of the Google-powered search engine. Google's success is, for the most part, traced back to the assumed high quality of its search results. Though Google is classified as being superior to its competitors, the question is if and how far this is really true. If the quality of the search results is the most decisive factor for users when choosing a search engine, then there is an urgent need to find ways to measure it, at least approximately, in an objective way. The aim of this research is to compare the retrieval effectiveness of Google and some of its prominent competitors. The intension was on the one hand, to measure the quality of the search results which give users hands-on hints as to which of the search engine is the best to choose when in need of information and, on the other hand to contribute to the theoretical question of how to measure the quality of web information retrieval systems.

The aim of this investigation is to compare the retrieval effectiveness of Google and some of its competitors, wherein the quality of the search results serves as the benchmark for the performance appraisal. What is meant by the quality of search results? If there is a difference, is it relevant? If there is a difference among the top rated search engines in retrieving relevant results? These are the most essential area of research of web search services. When using search engines the descriptions of results are the first things searchers see, and the results are what they get in the end. Why is it important to differ between the results themselves and the presentation of results on search engine result pages? Generally, in a first step, users decide with the help of the result presentations which hits deserve further investigation. Only those
that seem to be relevant have a chance of being selected. Hence the presentations of the results predetermine the user choices of assumed relevant items, regardless of the real quality of the Web pages. If the quality estimations given by the result lists correspond with the true quality of the results, there is no problem; on the contrary, the overview enables users to differentiate rapidly between good and bad results. Users could waste time trying to detect whether would-be relevant pages serve their information need. In a worst case scenario users would miss relevant hits, if the result presentation is inadequate. Again there is the possibility of missing required information, if search user confined their search within a particular search engine. Therefore, it is necessary to evaluate the performance of the most popular search engines in retrieving relevant internet resources. This study will also help to differentiate that which search engine is most reliable to meet the needs of the web searchers.

1.8 Chapter Plan


The first chapter provides an introduction to the study elaborating the problem, objectives and justification of the study, scope of the study, limitation, hypotheses of the study, and methodology.
The second chapter discusses the review of the related literature in connection with the present study.

The chapter three deals with the concepts of World Wide Web and the statistical growth of web information. It has also been discussed in details about the traditional and web information retrieval systems in this chapter.

In chapter four, it has been discussed about the search engine, the history and development of search engines, and the quantitative study of search engine. An year over year list of world most used search engines has also been presented in this chapter. A details overview of the five most used search engines have been presented in this chapter.

In chapter five, a comprehensive discussion on the searching strategies of selected search engines has been made. A detail discussion on the existing facilities of search engines has been made in terms of Interface, Search Features, Limit options, Results Display, Case sensitivity, and other features.

In chapter six, a detail evaluation of the performance of the selected search engines have been made on the basis of data analysis which were collected from two different studies made in different time frame. In this chapter, the Coverage, Relevancy, Time factor, Stability, and Uniqueness of search results has been analyzed by using the SPSS software.
Chapter seven discusses the findings of the present study systematically and provides overall conclusion and suggestions to be adopted for better search results and scope for future research in this field.

The Bibliography and References along with the Annexures are appended at the end of the work.