EVALUATING AND COMPARING SEARCH ENGINES IN RETRIEVING TEXT INFORMATION FROM THE WEB

ZEMICHAEL FESAHATSION WELDEGHEBRIEL

Assignment presented in partial fulfilment of the requirements for the degree of Master of Information and Knowledge Management at the University of Stellenbosch.

SUPERVISOR: Dr. M. S. VAN DER WALT

APRIL 2004

DECLARATION

"I, the undersigned, hereby declare that the work contained in this assignment is my own original work and that I have not previously in its entirety or in part submitted it at any University for a degree.

ABSTRACT

With the introduction of the Internet and the World Wide Web (www), information can be easily accessed and retrieved from the web using information retrieval systems such as web search engines or simply search engines. There are a number of search engines that have been developed to provide access to the resources available on the web and to help users in retrieving relevant information from the web. In particular, they are essential for finding text information on the web for academic purposes. But, how effective and efficient are those search engines in retrieving the most relevant text information from the web? Which of the search engines are more effective and efficient? So, this study was conducted to see how effective and efficient search engines are and to see which search engines are most effective and efficient in retrieving the required text information from the web. It is very important to know the most effective and efficient search engines because such search engines can be used to retrieve a higher number of the most relevant text web pages with minimum time and effort.

The study was based on nine major search engines, four search queries and relevancy judgments as relevant/partly-relevant/non-relevant. Precision and recall were calculated based on the experimental or test results and these were used as basis for the statistical evaluation and comparisons of the retrieval effectiveness of the nine search engines. Duplicated items and broken links were also recorded and examined separately and were used as an additional measure of search engine effectiveness. A response time was also recorded and used as a base for the statistical evaluation and comparisons of the retrieval efficiency of the nine search engines.

Additionally, since search engines involve indexing and searching in the information retrieval processes from the web, this study first discusses, from the theoretical point of view, how the indexing and searching processes are performed in an information retrieval environment. It also discusses the influences of indexing and searching processes on the effectiveness and efficiency of information retrieval systems in general and search engines in particular in retrieving the most relevant text information from the web.

OPSOMMING

Met die koms van die Internet en die Wêreldwye Web (www) is inligting maklik bekombaar. Dit kan herwin word deur gebruik te maak van inligtingherwinningsisteme soos soekenjins. Daar is 'n hele aantal sulke soekenjins wat ontwikkel is om toegang te verleen tot die hulpbronne beskikbaar op die web en om gebruikers te help om relevante inligting vanaf die web in te win. Dit is veral noodsaaklik vir die verkryging van teksinligting vir akademiese doeleindes. Maar hoe effektief en doelmatig is die soekenjins in die herwinning van die mees relevante teksinligting vanaf die web? Watter van die soekenjins is die effektiefste? Hierdie studie is onderneem om te kyk watter soekenjins die effektiefste en doelmatigste is in die herwinning van die nodige teksinligting. Dit is belangrik om te weet watter soekenjin die effektiefste is want so 'n enjin kan gebruik word om 'n hoër getal van die mees relevante tekswebblaaie met die minimum van tyd en moeite te herwin.

Heirdie studie is baseer op die sewe hoofsoekenjins, vier soektogte, en toepasliksheidsoordele soos relevant /gedeeltelik relevant/ en nie- relevant. Presiesheid en herwinningsvermoë is bereken baseer op die eksperimente en toetsresultate en dit is gebruik as basis vir statistiese evaluasie en vergelyking van die herwinningseffektiwiteit van die nege soekenjins. Gedupliseerde items en gebreekte skakels is ook aangeteken en apart ondersoek en is gebruik as bykomende maatstaf van effektiwiteit. Die reaksietyd is ook aangeteken en is gebruik as basis vir statistiese evaluasie en die vergelyking van die herwinningseffektiwiteit van die nege soekenjins.

Aangesien soekenjins betrokke is by indeksering en soekprosesse, bespreek hierdie studie eers uit 'n teoretiese oogpunt, hoe indeksering en soekprosesse uitgevoer word in 'n inligtingherwinningsomgewing. Die invloed van indeksering en soekprosesse op die doeltreffendheid van herwinningsisteme in die algemeen en veral van soekenjins in die herwinning van die mees relevante teksinligting vanaf die web, word ook bespreek.

DEDICATION

This paper is dedicated with love to my father, Fesahatsion Weldeghebriel, and my mother, Hiwet Beyene, for their encouragement, inspiration and endless love.

ACKNOWLEDGEMENT

This research assignment has benefited from the inputs of many people. Although my gratitude is to all, I would like to mention some of them. First and foremost, my deepest appreciation and gratitude goes to my supervisor, Dr. M. S. Van der Walt, for his expert guidance, continued advice and fruitful comments on my work as well as for his friendly attitude.

My special acknowledgement of gratitude also goes to Siham Mukhtar for her valuable and constructive suggestions, starting from the day of formulating the research problems until the end, as well as for helping in the electronic formatting of my research paper.

My special thanks and gratitude go to Zerai Ghebretensae for doing the statistical analysis as well as to Hermon Ogbamichael for his proof reading of Chapter Two and Three of my research paper. A special acknowledgement of gratitude also goes to Melanie Bailey for editing my research paper as well as for translating the abstract into Afrikaans.

My special thanks go to the government of the State of Eritrea for the scholarship that covered all my academic and living expenses. I would like also to thank to all my family members for their prayers, encouragement and endless love. Finally, I wish to extend my appreciation to all my friends here in Stellenbosch and all over for their motivation and intellectual stimulation.

My acknowledgement of gratitude expressed here goes beyond the limits of the brief way in which they are all noted. I am very grateful.

TABLE OF CONTENTS

Declaration ii
Abstract iii
Opsomming iv
Dedication v
Acknowledgement vi
List of Tables x
Glossary and List of Abbreviations xi

Chapter One: Introduction 1
1.1. Background and Problem Statement 1
1.2. Motivation of the Study 2
1.3. Objectives of the Research 2
1.4. Research Methodology 3
1.5. Organization of the Paper 3
Chapter Two: Information Retrieval and Indexing 5
2.1. Introduction 5
2.2. Information Retrieval (IR) 5
2.3. Information Retrieval Systems (IRS) 6
2.4. Functions of Information Retrieval Systems 6
2.5. Indexing 7
2.5.1. Definition of Indexing 7
2.5.2. Indexing Processes 8
2.5.3. Types of Indexing Processes 10
2.5.3.1. Manual Indexing 10
2.5.3.2. Automatic Indexing 10
2.6. Indexing Languages 12
2.6.1. Controlled Indexing Languages 13
2.6.2. Natural Indexing Languages 13
2.7. Exhaustivity and Specificity of Indexing Languages 14
2.7.1. Indexing Exhaustivity 14
2.7.2. Indexing Specificity 15
2.8. Precision and Recall in IRS 15

2.8.1. Precision 15
2.8.2. Recall 16
2.9. Summary 17
Chapter Three: Searching and Information Retrieval Tools 19
3.1. Introduction 19
3.2. Searching 19
3.2.1. Search Processes 20
3.2.2. Search Strategies 21
3.2.3. Features of Searching 22
3.2.3.1. Single Word Searching 22
3.2.3.2. Boolean Search Statements 23
3.2.3.3. Phrase Searching 24
3.3. Influences of Searching in Information Retrieval Effectiveness 25
3.4. Information Retrieval Tools 26
3.4.1. Directories 26
3.4.2. Search Engines 27
3.4.2.1. Indexing and Search Engines 28
3.4.2.2. Searching and Search Engines 29
3.5. Major Search Engines 29
3.6. Summary 31
Chapter Four: Methodology for Evaluating and Comparing Search Engines 33
4.1. Introduction 33
4.2. Reasons for Evaluating IRS 33
4.3. Effectiveness and Efficiency in IRS 34
4.4. Types of Search Engine Evaluation Approaches 34
4.5. Previous Studies 35
4.6. Features of Search Engine Evaluation 37
4.7. Search Queries and Relevance Judgement 38
4.7.1. Search Queries 38
4.7.2. Relevance Judgment 39
4.8. Evaluation Criteria or Measurements 40
4.9. Summary 44
Chapter Five: Evaluation and Comparison Results of Search Engines 46
5.1. Introduction 46

5.2. Discussion of Experimental Results 46
5.3. Retrieval Effectiveness 47
5.3.1. Precision Results 47
5.3.2. Recall Results 48
5.3.3. Over All Effectiveness: Precision and Recall 49
5.4. Effectiveness Results in Eliminating Duplicated Items and Broken Links 50
5.4.1. Effectiveness in Eliminating Duplicated Items (DI) 50
5.4.2. Effectiveness in Eliminating Broken Links (BL) 51
5.5. Retrieval Efficiency 52
5.6. Effective and Efficient Search Engines 53
5.7. Summary 53
Chapter Six: Conclusions and Recommendations 55
6.1. Conclusions 55
6.2. Recommendations 60

REFERENCES 62

LIST OF TABLES

Table 4.1: Desirable Features of Web Search Evaluation 37
Table 5.1: Total Number of Relevant (R) and Non-Relevant (NR) Items 46
Table 5.2: Total Number of Duplicated Items (DI) and Broken Links (BI) 47
Table 5.3: Precision of Search Engines for each Query (in %) 47
Table 5.4: ANOVA for Precision of Search Engines 48
Table 5.5: Recall of Search Engines for each Query (in %) 48
Table 5.6: ANOVA for Recall of Search Engines 49
Table 5.7: Average Precision and Recall of Search Engines (in %) 49
Table 5.8: ANOVA for the Average Precision and Recall 50
Table 5.9: Total Number of Duplicate Items (DI) for each Query (in %) 50
Table 5.10: ANOVA for the Duplicated Items 51
Table 5.11: Total Number of Broken Links (BI) for each Query (in %) 51
Table 5.12: ANOVA for Broken Links 52
Table 5.13: Response Time (in Seconds) 52
Table 5.14: ANOVA for the Response Time 53

GLOSSARY AND LIST OF ABBREVIATIONS

1. GLOSSARY

The following words are used throughout the paper and are defined as follows:

- **Database** - is a collection of related records.
- **Document** - is any physical form of recorded information (Foskett, c1999:3).
- **Index Terms** - are terms used to represent a document or an item in a database.
- **Indexer** - is a person who indexes documents or items.
- **Information** - is processed data and facts that have been organized and communicated in a coherent and meaningful manner in textual, numerical, graphical or other form.
- **Information Need** - is an explanation of the information one would like to receive from a search (Gordon & Pathak, 1999:146).
- **Item** - is used to represent a unit of document contents, information, words and so on.
- **Keyword** - is a word or words used to represent documents or items or concepts during indexing and search processes.
- **Needed information** - is information that is useful to the user or searcher.
- **Non-Relevant Items** - are items that do not provide any useful or required information.
- **Relevant Item** - is an item containing the required information
- **Requested Information** - is information that one would like to receive from a search.
- **Required Information** - is information that relates to a user's need.
- **Searcher** - is a person who searches information for his own purposes or those of others.
- **Search Item** - are keywords or a combination of keywords used to search the required information.
- **Search Query/Query** - is a description of the needed information that a user or searcher uses to communicate with the system to retrieve the required information.
- **Search Statements** - are keywords or combination of words that describes the needed information is entered into the system to retrieve the needed information.
- **User** - is a person who searches information for his own purposes.

Note: In this study, the meaning of information is limited to text only. Moreover, document & item; needed information, requested information and required information; search item, search statements and search query/query; user(s) and searcher(s) are used interchangeably unless they are identified by the context.

2. LIST OF ABBREVIATIONS

ANOVA	Analysis of Variance
BL	Broken Link(s)
DCV	Document Cut-off Values
DI	Duplicated Item(s)
IR	Information Retrieval
IRS	Information Retrieval System(s)
LSD	Fishers Least Significant Difference(s)
NR	Non-Relevant
PR	Partly-Relevant
R	Relevant
TREC	Text Retrieval Evaluation Conferences

Chapter One

Introduction

1.1. Background and Problem Statement

In the modern age, which is the information age, information is available on the Internet, which is a collection of interlinked computer networks that exchange information according to some agreed protocols (Ackermann & Hartman, c1999:2), or on the World Wide Web (www), which is a large collection of information that is connected or linked in a sort of web (Ackermann & Hartman, c1999:1). The required information can easily be accessed and retrieved from the web using information retrieval systems, which are systems that are developed to store, retrieve, organize and maintain information on the web. In other words, users search and retrieve information from the web using information retrieval tools.

Information retrieval tools or mechanisms that are used to retrieve information from the web are web directories and search engines. Web directories are subject catalogues, which are collections of Internet and web resources arranged in categories (Ackermann & Hartman, c1999:409) and can be used to retrieve information from the web in the case of general and single faceted topics. Web search engines are retrieval services consisting of databases containing mainly the resources available on the web (Chowdhury, c1999:402) and can be used to retrieve information from the web in the case of very specific and multifaceted topics.

There are a number of search engines that have been developed to provide access to the resources that are available on the web. All of them index and store each of the web documents on their databases. They are full text databases. However, each one of them has a different way of determining which search items are most relevant to the users' requests (Ackermann & Hartman, c1999:126). But, all of them retrieve the requested information from the web by matching the users' search queries with the indexed terms. So, the question here is how effective and efficient are those search engines in retrieving the most relevant text information from the web by matching the search queries with the indexed terms? Which search engines are more effective and efficient in retrieving the requested information from the web?

Therefore, it is important to evaluate and compare those search engines to see whether they retrieve the requested information from the web in an effective and efficient manner by matching the search queries with the indexed terms or not. To do that, it is very important to identify evaluation criteria that could be used in the evaluation and comparison processes. In addition, since search engines involve indexing and searching processes in the information retrieval processes from the web, it is important to see, from the theoretical point of view, how the indexing and searching processes are performed in information retrieval systems as well as their influence on the effectiveness and efficiency of those systems in general and search engines in particular in retrieving the most relevant information from the web. However, this study mainly deals with the evaluation and comparison of search engines in retrieving text information from the web. It doesn't include directories in the evaluation and comparison processes.

1.2. Motivation of the Study

In the information age, information is valuable for competitiveness and users search information from the web for different purposes. Students and researchers especially retrieve text information from the web for academic purposes using information retrieval tools. So, the motivation of this study is based on the fact that identifying the most effective and efficient web search engines for retrieving the most relevant text information are important at academic level. Hence, this study will help students and researchers to identify and use the recommended search engines to retrieve the most required information from the web more efficiently and effectively. It might also serve as a motivation for search engine providers to upgrade their search engine standards.

1.3. Objectives of the Research

The specific objectives of this study are:

- To see how the indexing and searching processes are performed and their influence in the effectiveness and efficiency of information retrieval systems in particular search engines, from the theoretical point of view.
- To identify the criteria or measurements that have to be used in evaluating and comparing search engine performances that are effectiveness and efficiency.
- To determine the effectiveness and efficiency of those search engines in retrieving the most relevant text information from the web.

- To determine which search engines are more effective and efficient in retrieving the most relevant text information from the web.
- To recommend the best search engines to students and researchers.

1.4. Research Methodology

First and foremost, detailed literature studies involving current literature on information retrieval subjects was conducted to produce ideas and concepts from various sources such as books, Internet sources, journals, conference proceedings and so on. After that, a specific set of evaluation criteria or measurements were developed to test the effectiveness and efficiency of web search engines in retrieving relevant information from the web.

After developing these evaluation criteria, four search terms or queries were developed from the field of Information and Knowledge Management. Those four search terms or queries were used to test the various search engines. The test was conducted for twelve consecutive days from 7:00 pm -12:00 am South African time. Each of the search terms or queries was entered once in all search engines to retrieve the required text information from the web. The first twenty of the retrieved items were taken into consideration and judged for relevancy based on the specified set of relevance judgement criteria. Then, the required data was collected.

After completion of the test and collection of the required data, the data was integrated and analysed based on the statistical analysis, which are analysis of variance (ANOVA) and Fisher Least Significant Difference (LSD) method. Lastly, the process of evaluation and comparison were done based on the statistical results.

1.5. Organization of the Paper

In chapter two, first, the detailed concepts of information retrieval and information retrieval systems on the web are discussed. Then, the processes of text document representations, which are indexing, with their detailed and related concepts as well as their influences on the effectiveness of information retrieval systems are discussed. Lastly, the two most important parameters for measuring the performance of any information retrieval systems, which are precision and recall, are discussed in this chapter.

Next, the detailed and related concepts of searching and their influence on information retrieval systems, in particular on search engines, are discussed in chapter three. Besides, the information retrieval tools that are used by users and searchers to retrieve information from the web resources are identified and their detailed concepts with particular emphasises on search engines are discussed.

In chapter four, the related concepts in the evaluation and comparisons of information retrieval systems, in particular search engines, are discussed in detail. The required search queries and the relevancy judgement criteria are identified. Especially, the evaluation or measurement criteria are identified in this chapter. In general, this chapter deals with the methodologies that should be followed in evaluating the performances of search engines in retrieving relevant text information from the web. Then, the next chapter, which is chapter five, deals with the experimental results as well as with the evaluation and comparison results of the nine major web search engines in retrieving the requested text information from the web. In other words, the more effective and efficient search engines are identified in this chapter.

Finally, chapter six concludes the study and some recommendations are made in this chapter.

Chapter Two

Information Retrieval and Indexing

2.1. Introduction

In the information age, documents are stored and recorded on the web and they are available to end users or searchers at any time. In other words, individuals and group of users have access to the web databases and retrieve the required information from the recorded web databases. Especially students and researchers retrieve text documents from the web for academic purposes. Those text documents are available on the web in a way that is suitable for retrievals. Thus, text documents are represented with keywords and stored on the web for the purpose of future retrieval. However, the question here is how documents in general and text documents in particular are represented in the web databases suitable for future retrievals? And how does representation of these text documents influence the retrieval effectiveness from the web of information retrieval systems from the web?

Therefore, this chapter tries to discuss the processes of text document representations, which is indexing, with their detailed and related concepts as well as their influence on the effectiveness of information retrieval systems. But, in order to have a good understanding, first, let's discuss in detail the concepts of information retrieval and information retrieval systems on the web.

2.2. Information Retrieval (IR)

Information retrieval deals with the retrieval of information from previously stored or recorded information or text documents. Chowdhury (c1999:1) points out that information retrieval presupposes that there are some documents or records containing information that have been organized in an order suitable for easy retrieval. The idea is that, in order to retrieve the required information so easily, the documents that contain the information being retrieved must be organized and stored in a searchable manner. Information cannot be retrieved from unorganized documents so efficiently and effectively, or may be not at all.

The main task of information retrieval is to extract relevant documents from a large collection of documents in response to a user's queries (Sparck Jones & Willett, c1997:317). it is also largely concerned with the designing, implementing and

maintaining of effective and efficient information retrieval systems for practical use (Chowdhury, c1999:372), that is, to retrieve the required information more efficiently and effectively. Thus, information retrieval systems play a great role in storing, organizing and retrieving the required information so easily.

2.3. Information Retrieval Systems (IRS)

Information retrieval systems (IRS) are systems that are capable of storage, retrieval and maintenance of information (Kowalski & Maybury, c2000:2). Chowdhury (c1999:2) also states that an IRS is a system that is designed to retrieve the documents or information required by the user community. In particular, IRS can be designed to retrieve actual texts that supposedly will satisfy the user's requirements. These systems are called full text retrieval systems, because they deal with the retrieval of the actual text documents. In general, information retrieval systems, according to Sparck Jones & Willet (c1997:98), are the complete organization for obtaining, storing, and making available information to users.

The basic task of an information retrieval system is to retrieve documents or texts with information content that is relevant to a user's information need (Sparck Jones & Willett, c1997:1). So, the aim of an IRS is to retrieve documents or information in response to a user's request in such a way that the contents of the documents or the information are relevant to the user's requirement (Chowdhury, c1999:333). Chowdhury (c1999:70) also further states that the objectives of any IRS are to retrieve all the documents that are relevant to the query and simultaneously withhold all those that are not relevant to the query. That is why one of the major functions of an IRS, as listed in 2.4, is to match the contents of the documents with the user's queries in order to retrieve the information most relevant to the user.

2.4. Functions of Information Retrieval Systems

Information retrieval systems are interacting systems that carry out the process of inputs to produce outputs. In other words, a user inputs his information need or queries and the IRS executes the queries to retrieve the required information, which is the output, by matching the user's search query with the documents in the databases. So, information retrieval systems can have various functions.

According to Chowdhury (c1999:3), some of the major functions of IRS are:

- To identify the information sources relevant to the areas of interest of the target users' community
- To analyse the content of the sources or documents
- To represent the analysed sources in a way that will be suitable for matching users' queries
- To analyse users' queries and to represent them in the form that will be suitable for matching with the database
- To match the search statement with the stored database
- To retrieve the information that is relevant
- To make the necessary adjustments in the system based on feedbacks from the user

In short, the functions of any IRS are to identify, analyse and represent documents in their databases, and then retrieve the requested information by matching the users' query with document contents in the databases.

So, as is implied in the above, one of the most important aspects of IRS is concerned with how the text documents in the systems or databases are represented. In other words, the critical aspects of an IRS that determine its effectiveness are how to represent concepts in a document (Kowalski & Maybury, c2000:51). This leads to the concept of indexing. Hence, the following topic discusses the issue of text document representation, which is indexing. But, first let's define what indexing means.

2.5. Indexing

2.5.1. Definition of Indexing

Indexing is the process of constructing document surrogates by assigning identifiers to text items (Chowdhury, c1999:56). Rowley & Farrow (c2000:125) says that indexing is the assigning of terms or codes from a list to specific documents on the basis of subject analysis or interpretation of concepts in the documents. Sparck Jones & Willett (c1997:1) further states that indexing refers to the way documents are represented for retrieval purposes. Therefore, indexing can be generally defined as the process of identifying and assigning of descriptors or keywords or terms or codes through subject or concept analysis of contents of the documents being indexed for the purpose of future retrieval. In short, indexing is the process of representing a text document with an indexing term for retrieval purposes.

2.5.2. Indexing Processes

Since indexing is the process of document representation by indexing terms through subject analysis for future retrieval, it is, therefore, the first and most important step in operating an IRS, especially in web search engines. The main objective of any indexing process is to fully represent a document or the content of a document with an appropriate indexing language so that the required information will be retrieved later on more efficiently and effectively. Chowdhury (c1999:88) states that the objective of good indexing is to isolate all the relevant documents or content of documents in a collection from the other documents in the same collection that do not discuss the desired topic. That is, one has to choose words for indexing that can differentiate a given document or document contents from all the others in the same collection.

So, in order to differentiate or isolate documents or document contents through indexing terms or words, and to choose those indexing terms or words that can best represent the document or contents of the document, the first task in the process of indexing is to determine exactly what the given document is all about. That is, the determination of the "aboutness" of the document. The "aboutness" of a document or document contents can also be determined through the processes of content or subject analysis of the documents being indexed.

Subject analysis deals with the conceptual analysis and the translation of this conceptual analysis into the conceptual framework of the indexing language (Taylor, c1999:132). Chowdhury (c1999:57) has also explains that subject analysis means the analysis of the thought content embodied in a document. That means, subject analysis is the process of determining the "aboutness" of document content and it includes the task related to the analysis as well as organization and storage of information (Chowdhury, c1999:3). It is, therefore, the most important task in the process of indexing. The reason is that subject analysis of the contents of documents helps to find the appropriate index terms that can best describe or represent the document that is being analysed. Then, these index terms can be likely used by the user or searcher in retrieving the required documents.

So, the basic task of indexing processes is to analyse the entire document in order to identify the content identifiers or keywords that can represent the document being analysed. Thus, indexing involves the analysis of the contents of the documents being

indexed and is the representation of document contents by some content identifiers or indexing languages for future retrieval. Therefore, indexing processes deal with the processes of conceptual or subject analysis of the contents of text documents and the translation of those conceptual or subject analyses into particular indexing languages.

In general, indexing processes include:

- Conceptual or subject analysis of the contents of the documents being indexed
- Determining the "aboutness" of the document contents
- Identifying descriptors or keywords
- Assigning indexing terms to the contents of the document
- Organizing, recording and storing in databases for future retrieval

The main problem in the indexing process is to find the content identifiers or indexing terms that can fully represent the document being indexed so that they can be matched later on with the users' search terms. This means, as Chowdhury (c1999:56) pointed out that the basic problem involved in the process of indexing documents is the choice of appropriate keywords or descriptors to represent the document. Those keywords must also be likely to be chosen by the users to retrieve the document. Unfortunately, there is a lack of logical and consistent procedures in subject analysis. Since there is no any mechanism to control the subject analysis of the document contents, different indexers may analyse the contents of a given document differently and as a result they may choose different index terms to represent the contents of the document.

Although a number of vocabulary control tools such as thesaurus, classaurus, etc. have been developed in order to choose appropriate keywords that represent documents or document contents during the indexing processes, they still require intellectual capabilities. Consequently, those tools are found inefficient (Chowdhury, c1999:57). In addition, the modern computer aided content analysis, which is based on the statistical analysis in the process of indexing, is found to be inefficient. The reason for this is that they choose the index terms or keywords based on the statistical calculations of the occurrences of keywords in the document. In other words, the higher the occurrence of the word or term in a given document the more significant the word or term is considered, which is not necessarily correct.

Therefore, the main challenge in indexing processes is to choose indexing terms that can fully and accurately represent the documents being indexed so that the same term is likely to be chosen by the searchers during the search processes. So, the process of indexing influences the effectiveness of IRS, because documents are retrieved on the basis of the correspondence between the search terms expressed in a query and the index terms of the document.

2.5.3. Types of Indexing Processes

2.5.3.1. Manual Indexing

Manual indexing is an indexing process in which the content identifiers or terms or keywords are selected manually. Here again, the first task is to do subject analysis to determine exactly the "aboutness" of the given documents that are being indexed. In manual indexing processes, subject analysis of the documents for identifying the index terms or keywords is done manually. This is based on a detailed analysis and interpretation of the text of documents (Sparck Jones & Willett, c1997:305). Manual indexing is a good way to choose appropriate indexing terms to represent or describe a given document, because a good indexing system differentiates all the relevant document contents from others in the same collection that do not discuss the required topics. Thus, manual indexing makes it possible to isolate or discriminate between the index terms of the relevant documents from others.

However, manual indexing processes pose their own challenges. High intellectual ability is needed and more than one indexer performs the indexing processes. Moreover, too much time is wasted in the process of indexing. It is also very difficult to be consistent although the indexing can be carried out accurately and at the right level of detail (Chowdhury, c1999:83). The reason is that different indexers analyse and index a given document differently, because they have different intellectual capabilities in subject analysis. Due to lack of consistency and the limitation of indexers' intellectual capabilities, manual indexing greatly affects the effectiveness of information retrieval. Hence, it affects the effectiveness and efficiency of information retrieval from the web.

2.5.3.2. Automatic Indexing

Automatic indexing is an indexing process in which the content descriptors or keywords are identified with the help of modern technology. Salton (as quoted in Chowdhury, c1999:87) states that automatic indexing is the assignment of the content identifiers

which is carried out with the aid of modern computing equipment. Therefore, in automatic indexing, the subject analysis of the contents of a given document is carried out by the mechanical analysis of the words in the given documents. In other words, some statistical measurement, such as total word frequency, frequency distribution and so on, are used to select keywords or index terms that can best represent or describe the contents of the documents. Thus, automatic indexing uses various sorts of frequency criteria to select words from the natural language of the given document texts that are being indexed. In most cases, if a word occurs frequently it will serve as a content identifier or an index term of the document being indexed. In addition to word frequency other measures such as the position of the word, e.g. in the title and headings, are also used in automatic indexing.

Automatic indexing processes might have some advantages. Salton (as quoted in Chowdhury, c1999:87) as well as Kowalski & Maybury (c2000:60) points out that some of the advantages of automatic indexing are:

- The level of consistency can be maintained in indexing
- The cost of index entries can be reduced in the long run
- Indexing time can be reduced
- Better retrieval effectiveness can be achieved and;
- Above all, the lack of the human expertise can be overcome by the intelligent use of the free text vocabularies.

However, it is very difficult to say that automatic indexing processes are absolutely effective and efficient in indexing the contents of the given documents and in enhancing retrieval effectiveness. First of all, automatic indexing uses frequency criteria to select keywords or index terms to represent the contents of the document. In other words, words which occur often will be chosen as content identifiers or index terms of the documents being indexed, although they are not necessarily the most significant. They may retrieve more information during the retrieval processes, but, they may lose the relevance of the information that is being retrieved, because high occurrence does not always mean that the word adequately represents the contents of the document.

For instance, the word “information” appears in many documents and will probably be selected as an index term in automatic indexing. But, this will not be a good indexing term because it can retrieve the whole collection in which most of the retrieved

information will be non-relevant to the user's requirement. Therefore, this is one of the drawbacks of automatic indexing.

Secondly, it is clear that good indexing must isolate all the relevant documents in a collection from the others in the same collection that do not discuss the desired topic. But, automatic indexing does not always do this. The reason is that a word in automatic indexing processes is regarded as a significant word if it appears several times in the contents of a document and serves as an index term. This indicates that in automatic indexing, subject analysis is not carried out accurately and in great detail. Consequently, it is not possible to assign a good indexing term that can make documents as different as possible, and this affects the retrieval effectiveness of information retrieval processes. The idea is that while a greater separation between documents enhances retrieval effectiveness, less separation will depress retrieval effectiveness (Rijsbergen, c1999:15). For instance, the word "information", as mentioned above, could not serve as an index term because it doesn't separate or discriminate the relevant documents from the others in the same collection.

Therefore, automatic indexing processes, in turn, influence the effectiveness and efficiency of information retrieval processes in retrieving the required information for the user from the web.

2.6. Indexing Languages

Indexing languages can be defined as terms or codes that might be used as access points in an index (Rowley & Farrow, c2000:125). In other words, indexing languages are terms or identifiers or descriptors that describe or represent the content of documents that are being indexed. Indexing languages are languages that are used to describe documents in databases during the indexing process. The elements of indexing languages are index terms, which may be derived from the text of the document to be indexed or might be arrived at independently (Rijsbergen, c1999:13).

According to Sparck Jones & Willett (c1997:305), indexing languages are designed to meet two requirements:

- First, to ensure that if the representations of documents and requests match the correspondence relevance relations hold

- Second, they are the means of achieving this and do not also allow or encourage matches where relevance relations don't hold

So, indexing languages are important in the processes of indexing, because information retrieval systems retrieve the user's information requirement by matching the user's queries with the indexed languages of the stored and indexed documents in the databases. Indexing languages include controlled indexing languages and natural indexing languages.

2.6.1. Controlled Indexing Languages

Controlled indexing languages are used for indexing documents from a list that are identified for assigning to specific documents. That is, as Chowdhury (c1999:119) states, controlled indexing languages are those in which both the terms that are used to represent subjects and the process whereby terms are assigned to particular documents are controlled by a person. They are approved indexing languages in a list to specific documents on the bases of subjective interpretation of the concepts in the documents (Rowley & Farrow, c2000:125-26). In this case, control is practised over which terms are used and the relationships between the terms are indicated. In addition, the searchers must choose their search terms from the controlled list. So, controlled indexing languages tend to ensure consistency in indexing processes and also tend to match the indexing languages of the indexers and searchers.

Controlled indexing languages are used in many information retrieval environments. Yet they are found inefficient for effective information retrieval processes from the web. In other words, controlled indexing languages seem to be more consistent, efficient and straightforward to searchers, but research has failed to prove this convincingly (Rowley & Farrow, c2000:127).

2.6.2. Natural Indexing Languages

Natural indexing languages are used for indexing documents by taking indexing terms or descriptors from the document being indexed. In this case, any terms in the document could be identified or selected for indexing terms (Rowley & Farrow, c2000:127). This means, in natural language indexing, any term that appears in the title, abstracts or text of a document record may be an index term. There is no mechanism to control the use of terms for such indexing. Similarly, the searchers are not expected to

use any controlled list of terms (Chowdhury, c1999:119 - 120). Thus, the full text of the document is taken into consideration in the processes of natural language indexing and the users are free to use their own search terms.

Natural indexing languages are used in many information retrieval environments, particularly in search engines. Yet, they are found inefficient for effective information retrieval processes from the web. The idea is that since natural indexing languages are using terms from the document being indexed, they retrieve more non-relevant information from the web. The reason is that searchers come without any knowledge of the indexed terms and of the enormous amount of information on the web that contains the same search terms in their documents.

Therefore, the above two types of indexing languages that are currently used in many information retrieval environments, particularly in search engines, significantly influence the retrieval of information from the web. However, the effectiveness of an indexing process and the overall retrieval system depend on two important factors or parameters, which are called exhaustivity and specificity of the indexing languages.

2.7. Exhaustivity and Specificity of Indexing Languages

2.7.1. Indexing Exhaustivity

Exhaustivity is the extent to which the different concepts in the document are indexed (Kowalski & Maybury, c2000:57). It is the degree to which the subject matter of a given document has been reflected through the index entries (Chowdhury, c1999:69). Foskett (c1999:23-24) states that exhaustivity is the extent to which the indexer analyses the given document to establish exactly what subject content the indexer has to specify. Similarly, Taylor (c1999:135) defines exhaustivity as the number of concepts that will be considered in the conceptual framework of the systems. In other words, exhaustivity refers to whether indexing is done on summarization level or on in-depth level, that is, indexing of all significant concepts in the document. Hence, if the indexing is in-depth, it increases the possibility of getting or retrieving a greater number of relevant items from the web. Otherwise, it reduces the possibility of retrieving a large amount of relevant information from the web.

2.7.2. Indexing Specificity

Specificity is the extent to which the indexing languages and search terms are precise in representing the given document in the processes of indexing and searching. It refers to the vocabulary of the system, and denotes the system by which the indexer specifies subject content when indexing (Rowley & Farrow, c2000:129). It is the ability of the indexing language to describe topics precisely (Rijsbergen, c1999:14). Keen & Digger (as quoted in Rijsbergen, c1999:14) further states that specificity is the level of precision with which a document is actually indexed. In other words, specificity refers to how broadly or how specifically the indexing terms are chosen in a given situation. The more specific the term, the better is the representation of the subject through the index entry (Chowdhury, c1999:70).

So, specificity indicates whether the index language describes topics precisely or not. It is the level of precision for which the document is accurately indexed. That means if the indexing terms are specific enough in describing the document, then they will increase the precision level in retrieving relevant information from the web. Otherwise, they will reduce the precision and increase the retrieval of non-relevant information from the web.

2.8. Precision and Recall in IRS

2.8.1. Precision

Precision is one of the most important parameters for measuring the performance of any information retrieval system. It refers to the proportion of the number of relevant documents retrieved by the system to the total number of documents retrieved (Chowdhury, c1999:70). In other words, precision means how a particular information retrieval system functions precisely in retrieving most of the items relevant to the user's requirement. It is clear that the objective of any information retrieval system is to retrieve relevant and only relevant items based on the users' queries and withhold the non-relevant items. So, precision does not only measure the accuracy of information retrieval systems in retrieving relevant items, but also it measures indirectly how far the information retrieval systems are able to withhold non relevant items in performing the information retrieval processes based on the user's request. Thus, it refers to the ability of an information retrieval system to withhold non-relevant items or documents. It can, therefore, be calculated as (Kowalski & Maybury, c2000:5; Ellis, c1996:7):

$$\Pr ecision = \frac{Number\_\mathrm{Re}trieved - \mathrm{Re}levant}{Number\_Total\_\mathrm{Re}trieved} \times 100\%$$

Where,

$Number\_\mathrm{Re}trieved\_\mathrm{Re}levant$ means the total number of relevant items retrieved by the query.

$Number\_Total\_\mathrm{Re}trieved$ means the total number of items retrieved by the query

E.g. if in a given query the system retrieves 80 items, out of which 50 are relevant and 30 are non relevant, then the precision is 0.625, which is 62.5%. This means that the system is 62.5% precise when retrieving the relevant documents or items in the given query.

However, the precision of an information retrieval system is affected by the indexing exhaustivity and specificity. That means, if the indexing is exhaustive, then it increases the possibility of getting or retrieving a greater amount of information from the web and reduces precision. On the other hand, if the indexing languages are specific, they reduce the number of retrieved items and increase precision.

2.8.2. Recall

Recall is another most important parameter for measuring the performance of any information retrieval system. It refers to the proportion of the number of relevant documents retrieved by the system to the number of relevant documents in the collection (Chowdhury, c1999:70). In other words, recall refers to the ability of an information retrieval system to retrieve all relevant items and this is the basic objective of any information retrieval system. It can be calculated as (Kowalski & Maybury, c2000:5; Ellis, c1996:7):

$$\mathrm{Re}call = \frac{Number\_\mathrm{Re}trieved - \mathrm{Re}levant}{Number\_Possible\_\mathrm{Re}levant} \times 100\%$$

Where,

$Number\_\mathrm{Re}trieved\_\mathrm{Re}levant$ means the total number of relevant items retrieved by the query.

$Number\_Possible\_\mathrm{Re}levant$ means the total number of relevant items in the database.

E.g. if there are 100 relevant items in a given database for the given query, out of which only 50 relevant items are retrieved, then the Recall is 0.5, which is 50%. This means that the information retrieval system has retrieved only 50% of the relevant items in the database.

Recall is also affected by the indexing exhaustivity and specificity of the indexing languages. That means, if the indexing is in-depth, then it increases the possibility of getting or retrieving a greater amount of information from the web and increases recall. On the other hand, if the indexing languages are specific, they reduce the number of retrieved items and reduce recall.

So, in general, precision and recall are the most important parameters for measuring the effectiveness of any information retrieval system. A good information retrieval system is one that can optimize both the precision and recall. And this means that it is necessary to balance the level of indexing exhaustivity and the specificity of the indexing languages during the indexing processes as well as the search terms during the searching processes.

2.9. Summary

Information retrieval deals with the retrieval of information from the previously stored or recorded information or text documents. Its basic task is to extract relevant information from the web based on the user's request. It also deals with designing of information retrieval systems in order to store, organize and retrieve the requested information. Information retrieval systems are the complete organization for obtaining, storing, and making available information to users.

The main functions of any IRS are to identify, analyse and represent documents in their databases, and then retrieve the requested information by matching the users' query with document contents in the databases. In other words, IRS retrieve the requested information based on the correspondence of the indexing terms with the search queries. An IRS is also concerned with how text documents in the systems or databases are represented or indexed.

Indexing is the process of representing a text document with an indexing term for retrieval purposes. In indexing processes, text documents are indexed through the process of subject analysis, which deals with the conceptual analysis and translating of

this conceptual analysis into the conceptual framework of the indexing. The two indexing processes are manual indexing and automatic indexing. In manual indexing, the subject analysis is done manually, whereas in automatic indexing, it is done automatically with the help of modern computing technologies.

In indexing processes, the indexing languages can be controlled indexing languages or natural indexing languages. Controlled indexing languages are approved indexing terms in a list, and these terms are assigned to specific documents on the bases of subjective interpretation of the concepts in the documents, whereas natural indexing languages are languages used for indexing documents by taking indexing terms directly from the document being indexed.

However, the effectiveness of the indexing process and the overall information retrieval system depend on two important factors or parameters, which are called exhaustivity and specificity. Indexing exhaustivity refers to whether indexing is done on summarization level or on in-depth level, that is, indexing of all significant concepts in the document, whereas specificity refers to the extent to which the indexing languages and search terms are precise in representing the given documents in the processes of indexing and searching respectively. But, in indexing processes, it is very difficult to represent a text document with exhaustive and specific indexing terms that can be chosen later on by the user. So, indexing processes greatly influence the effectiveness and efficiency of any information retrieval system.

The effectiveness of any information retrieval system can be measured by precision and recall. Precision refers to the proportion of the number of relevant documents retrieved by the system to the total number of documents retrieved, whereas recall refers to the proportion of the number of relevant documents retrieved by the system to the number of relevant documents in the collection. So, a good information retrieval system is one that can optimize both the precision and recall. This means that it is necessary to balance the level of indexing exhaustively and the specificity of the indexing languages during the indexing processes.

Chapter Three

Searching and Information Retrieval Tools

3.1. Introduction

Users visit the web resources frequently to retrieve information that is needed to perform their tasks more efficiently and effectively. Thus, they search for information from the web for different purposes. In particular, students and researchers retrieve text documents from the web resources for academic purposes using the different information retrieval tools. But, the questions here are:

- How is the searching process performed to retrieve information from the web resources?
- What search strategies should be used to retrieve the required and most relevant information more efficiently and effectively?
- How does searching influence the effectiveness of information retrieval systems?
- What are the information retrieval tools that are used by searchers to retrieve information from the web resources?
- How do these information retrieval tools work and perform their basic tasks?

Therefore, this chapter discusses these issues and their related concepts in detail. So, let's begin the discussion with the concept of searching.

3.2. Searching

Searching refers to the way the file is examined and the items in it are taken as related to a search query (Sparck Jones & Willett, c1997:1). In other words, searching refers to the retrieval of the needed information from its source databases based on the similarity between users' search queries and the documents in the web databases. So, searching involves users' search terms and supposes that there are items in the sources' databases that can match the users' search terms. It also assumes that there are retrieval systems in which the search terms are executed to retrieve the required information. In general, searching is concerned with search processes in which information is retrieved from a source based on users' search queries. So, the next topic deals with how the search process is performed.

3.2.1. Search Processes

In an information retrieval environment, search processes begin when a user or a searcher wants to search for some information that can answer for his/her information requirements or requests. That means a user or searcher comes with an information need for a particular purpose. So, the first and most important task of a user or searcher in a search process is to determine clearly his/her information needs or requirements as well as determining the "aboutness" of his/her information needs or requirements. Determining the "aboutness" of the information needs helps the user to develop adequate search statements or queries or search terms, which are short statements or terms that are used by a searcher when specifying search requirements (Rowley & Farrow, c2000:125).

In order to determine the "aboutness" of the information needs, the user or searcher, like the indexer, is required to do some conceptual analysis of the needed information. Conceptual analysis helps the searcher to formulate exhaustive and specific search statements or queries or search terms that can probably match the descriptors so that it can retrieve the most relevant items. In this case, the conceptual analysis and formulation of the search terms or queries are performed by the user or searcher. That means the user or searcher develops his/her own search statements based on his/her conceptual analysis of the required information during the actual search.

After formulating appropriate and adequate search terms, the user enters his/her queries into the selected information retrieval system to retrieve the intended or required information. The system then retrieves the requested information by matching the search terms with the index terms. Lastly, a user or a searcher reviews the retrieved items or hits for relevance because all the retrieved items are not necessarily relevant to the users' or searchers' requirement. This is due to the availability of many items in the web databases that contain the words of the search statements that are used by the searcher. Therefore, relevance judgement is the last step for a single round of search in the search processes from the web.

To summarise, search processes from the web include:

- Identifying user's information requirement that should be searched
- Conceptual analysis of the contents of the information to be searched
- Determining the "aboutness" of the contents of the information to be searched

- Formulating search statements or queries or search terms
- Executing the search statements or queries or search terms into the system
- Reviewing the retrieved items for relevance.

So, from the above discussion, it is clear that the main tasks of a searcher in the search process are to formulate appropriate search terms and perform the actual search, which is the retrieval of the required information. The reason is that search terms are the terms that will be entered into the system to retrieve the intended information from the web. Thus, search terms directly affect the retrieval effectiveness of the required information. That means if the search terms represent the concept adequately, then the system can retrieve the most relevant information. Otherwise, it will retrieve non-relevant information. So, it is very important to formulate appropriate search terms. To do that, a user or searcher must follow appropriate search strategies.

3.2.2. Search Strategies

Search strategies are sets of decisions and actions taken during the search processes (Rowley & Farrow, c2000:103). In other words, they are steps that a user or searcher takes during the search processes in formulating appropriate search statements or queries or search terms in such a way that those search statements can match with the indexing terms so that the system can retrieve the most relevant information efficiently and effectively. Chowdhury (c1999:158) further states that search strategies help the user or searcher to select the optimum path for searching the information in the databases. Thus, it helps to select appropriate sources, to formulate appropriate search terms and to select appropriate search techniques. So, in order to apply search strategies in their proper place during search processes, the searcher or user must take certain steps.

The most important and basic steps that a user has to take in formulating search strategies, i.e., formulating search queries, and viewing search results, are (Ackermann & Hartman, c1999:133):

1. Identify the important concepts of the search
2. Choose the keywords that describe these concepts
3. Determine whether there are synonymous, related terms, or other variations of the keywords that should be included
4. Determine which search features may apply

5. Create appropriate search statements
6. Determine which retrieval systems might be used
7. View the result and make a relevance judgement

So, from the above points, it is clear that one of the basic tasks of search strategies is to formulate appropriate search statements or queries or search terms that can make the searching processes successful in retrieving the most relevant information. But, in order to formulate appropriate search terms, a searcher needs once again to understand the various features or techniques of searching.

3.2.3. Features of Searching

There are many features of searching in which the search statements or search queries can be formulated. Some of the features that are widely used in formulating the search statements or queries are Single-word searching, Boolean searching, Phrase searching, Proximity, Truncation, Best-match searching and so on. These features are used in many information retrieval systems, particularly in search engines. However, this paper focuses on the first three features namely Single-word searching, Boolean searching, and Phrase searching.

3.2.3.1. Single Word Searching

Single-word searching is one of the features of searching in retrieving the required information from the web. In this case, the required information can be represented in a single word. In other words, the search statement contains only a single word, and this single word will be entered into the system during the search process to retrieve the required information.

Although a single word can be used for searching the information needed from the web, according to Strzalkowski (1995:397), simple or single word-based representation of the contents of the required information is usually inadequate since single words are rarely specific enough for accurate discrimination. Thus, it is usually difficult to express or represent a concept exhaustively and specifically with a single word. Hence, single word searching may not be effective in retrieving the most relevant items from the web. However, sometimes it can be effective and efficient in retrieving relevant items. For instance, if a user wants to get some information about Mostertsdrif, which is part of Stelenbosch, the search query can be formulated with a single word as “Mostertsdrif”. In

this case, the keyword, "Mostertsdrif", can retrieve the most required and relevant information from the web.

3.2.3.2. Boolean Search Statements

Boolean search statements are the most widely used in searching information from the web. They are an effective way of expressing search statements or queries. Thus, the search statements are connected with one operator or combination of operators. Boolean operators help to formulate appropriate search statements that can denote meaningful concepts by the breaking down of concepts into key concepts. In other words, they allow a user to logically relate multiple concepts together to define what information is needed (Kowalski & Maybury, c2000:29). They include the operators "AND", "OR" and "NOT". In some cases, the operators "AND" and "NOT" have to be replaced with "+" and "–" signs.

The operator "AND" is used mostly between the search terms when the user wants to narrow the search statement. It retrieves a small set of information and increases the precision. Whereas, the operator "OR" is used when the user wants to expand the search statements and looks for alternatives. It retrieves a large set of information and increases recall. The operator "NOT" is also used when a term needs to be excluded, and the information that was indexed with the excluded terms will not be retrieved. Hence, it retrieves a small set of information and increases precision.

For instance:

1. In order to ask for information on Intranets and Knowledge Management, the user may formulate the search statement as "Intranets and Knowledge Management". In this case, the search statement is narrowed and it can retrieve a small set of information. The retrieved items then are expected to discuss 'Intranets and Knowledge Management' as they can be considered as relevant items.
2. In order to ask for information on Intranets or Knowledge Management or on both', the user may formulate the search statement as "Intranets or Knowledge Management". In this case, the user or searcher widened his search statement, and the system retrieves a large set of information. The retrieved items then are expected to discuss the concepts of either intranets or Knowledge management or both as they can be considered as significant items.

3. In similar manner, in order to ask for information on Intranet and not on Knowledge Management, the user may formulate the search statement as "Intranets and not Knowledge Management". In this case, the user narrows his search statement and the retrieved items will be expected to discuss only the concept of Intranets, not where Knowledge Managements are also mentioned.

Although Boolean search statements are the most widely used in searching information from the web, they have their own drawbacks or limitations. First, it is very difficult to formulate an exact search statement by the combination of the operators "AND", "OR" and "NOT", especially when several terms are involved. In other words, it is very difficult to formulate nested Boolean search queries with the appropriate coordination of the search terms. Even if it can be used, either the search statement will become too narrow or too broad. In such cases, the user or searcher can miss the most relevant information. Second, Boolean searching identifies an item as relevant by finding out whether a given query term is present or not in a given record in the database. Thus, all retrieved items are considered to be of equal importance, that is, the retrieved items cannot be ranked in decreasing order of relevance (Chowdhury, c1999:161). Consequently, the user can miss the most relevant information.

3.2.3.3. Phrase Searching

A phrase is a string of words that must appear next to each other (Ackermann & Hartman, c1999:129). Therefore, a phrasal search statement is used when the user wants to see the words in the retrieved items next to each other in the same order as in the search statements. Ackermann & Hartman (c1999:129) further stated that phrasal searching is one of the most helpful search features and increases the chance of retrieving relevant information from the web, because, phrase searching is better in denoting the important concepts of the required information and it gives the required meaning. In other words, a phrase is better in representing concepts that are sought. In phrasal searching, it is required to put the search statements within double quotation marks in order to differentiate from word searching. Most search engines, for instance, require double quotation marks to perform the search as it is intended. For example, if a user wants to retrieve information that discusses the concept of information orientation, then the phrasal search statement will be written as "*Information Orientation*".

Although phrasal searching is one of the most helpful search features and increases the chance of retrieving relevant information from the web, it has its own drawbacks. The reason is that phrasal searching only considers the words next to each other in the same order as the search query to retrieve the requested information from the databases. Thus, only those items that contain the words like in the phrasal query are retrieved. In such cases, the phrasal searching can miss other relevant items that do not contain the words next to each other or in the same order as in the phrasal search statements, as for example in the phrases "*information and knowledge orientation*" or "*orientation towards information*".

3.3. Influences of Searching in Information Retrieval Effectiveness

During the search processes, a user or searcher tries to use the index terms that were used by the indexer in the process of indexing document contents. But, the searcher cannot approach with the index terms which were assigned to the documents during the indexing processes. In other words, the searchers do not come to the search processes with any knowledge of the specific document profiles, which are sets of search keys for the document (Rowley & Farrow, c2000:99) that they might think relevant (Rowley & Farrow, c2000:101). The reason is that the indexers, not the searchers, do the indexing processes either manually or automatically. So, searchers use their own search terms to search for the desired information during the search processes. In this case, it would be very difficult for the user to retrieve the required and relevant information, especially from the web. This is due the fact that an information retrieval system retrieves the requested information based on the correspondence between the search terms and the indexing terms. Hence, searching can influence the effectiveness and efficiency of information retrieval from the web databases.

Moreover, a searcher faces difficulties in approaching or formulating appropriate search expressions or queries. In other words, the main challenge for the user is that he/she is unable to formulate and adopt appropriate search strategy for searching for the desired information. The reason is that developing good search strategies requires knowledge about the nature and organization of the target information and also the exact needs of the searcher (Chowdhury, c1999:158). The idea is that many users do not begin with such a clear view of their information requirements that they can formulate a sharp search statement (Armstrong & Large, c2001:10), while the result of a search depends

heavily upon the correct understanding of the user's precise needs (Chowdhury, c1999:159).

In addition, users face difficulties in conceptualizing and analysing their information needs, and as a result they can develop inadequate search statements or queries, which greatly affect the retrieval of the required information from the source databases. For instance, as Kowalski & Maybury (c2000:166) points out, the length of search statements directly affect the ability of information retrieval systems to find relevant items. So, the lack of users' ability in formulating appropriate search statements or queries using appropriate search strategies and search features influences the information retrieval effectiveness from the web.

Therefore, from the above discussion, it can be deduced that search processes can greatly influence the retrieval effectiveness of any information retrieval systems from the web.

3.4. Information Retrieval Tools

As explained in the preceding chapter, the main task of information retrieval is to extract relevant documents from a large collection of documents in response to a user's query. It is mainly concerned with the designing, implementing and maintaining of effective and efficient information retrieval systems for practical use. Information retrieval also deals with retrieval tools or mechanisms to retrieve the required information, and thus, a number of information retrieval tools or mechanisms have been developed to retrieve the required information from the web databases. The well known information retrieval tools are directories and search engines.

3.4.1. Directories

Directories are subject catalogues, which are collections of internet and web resources arranged in categories (Ackermann & Hartman, c1999:409). Green (2000:125) also defines a directory as a predefined list of web sites, compiled by human editors and categorised according to subjects/topics. In a directory, subject headings are arranged in hierarchical lists and they are created and maintained by persons. In other words, people assign the subject headings to the records in the web databases and the web pages are indexed manually by the concerned persons. So, directories provide hierarchical menus of subjects that can be used to narrow a search (Armstrong & Large,

c2001:5) and searching is via menus of the added subject headings, that is, browsing by subjects or through keyword searching (Rowley & Farrow, c2000:312). Thus, the user navigates through the listing of topics or through keyword searching across the entire directory databases to retrieve the desired information.

Directories are effective in the retrieval of information from the web for general and single faceted topics because the subjects are arranged in a topical list and the user can simply follow the links to get the required information. They also contain fewer resources and so it is easier and less time-consuming to visit all the web pages. Besides, they rate, annotate, analyze, evaluate, and categorize the resources included, which helps the user to find the highest quality (Ackermann & Hartman, c1999:99). Thus, directories increase the probability of retrieving relevant information.

However, the major disadvantages in using some directories are that the hierarchical arrangement may be arbitrary. Infrequent updates as well as the subjectivity of rating and annotating resources by the indexers are also disadvantages. Above all, directories only give access to a fraction of the web (Armstrong & Large, c2001:5). In other words, when users search information using the directory, they have access only to those resources that are included in the directory, not the entire web (Ackermann & Hartman, c1999:103). Hence, the user may not retrieve a large set of items and thus may fail in retrieving some of the relevant information. Furthermore, the users' ability to assess the relevance of a document depends critically upon the metadata that is displayed about the document in the displays of the retrieved set (Rowley & Farrow, c2000:312). Thus, in such cases a user needs to use other information retrieval tools, which are search engines.

3.4.2. Search Engines

Search engines are retrieval mechanisms that perform the basic retrieval tasks, the acceptance of a query, a comparison of a query with each of the records in a database, and the production of a retrieval set as output (Rowley & Farrow, c2000:310). Salton & McGill (as quoted in Can, Nyray & Selvidik., 2003:2) also states that search engines are information retrieval systems, which are used to locate the web pages relevant to the users' queries. In other words, search engines are computer programmes that gather information about resources on the Internet by means of a robot, which is called a spider or crawler, and store this information in a database and make it accessible to

Internet users through a retrieval model that allows keyword searching (Van der Walt, c2000:182). Thus, the spiders or crawlers or robots gather new documents from the web, download them into the search engine databases and index them. The search engine then retrieves the required information from the database based on the users' search queries.

Thus, search engines are retrieval tools consisting of databases that contain mainly the resources available on the web (Chowdhury, c1999:402). Hence, the basic objective of any search engine is to retrieve relevant documents for the user from the search engine database based on the correspondence of the users' search queries with documents in the databases.

Search engines are effective tools to use when users or searchers are looking for very specific information or when the search topics have many facets or are multifaceted. They retrieve more and up-to-date information during the retrieval processes because they have access to those resources included in the entire web. Therefore, it is very easy to retrieve a large set of information from a given search query using search engines. However, it is also clear that not all retrieved items are relevant to the user's requirement. The reason is that search engines, like any other information retrieval system, are influenced by the processes of indexing and searching.

3.4.2.1. Indexing and Search Engines

Most search engines index the web pages through automatic indexing processes. That means the subject analysis for identifying indexing languages is performed automatically with the help of computer software, that is, robots or crawlers or spiders. In other words, the subject analysis of the contents of a given document is carried out by the mechanical analysis of the words in the given documents. Those computer software use some statistical measurements, such as total word frequency, frequency distribution and so on, to select keywords or index terms that can best represent or describe the contents of the documents. In most cases, words which occur frequently in the document will serve as content identifiers or index terms of the documents being indexed. They also use natural languages for representing or indexing the documents in their databases.

The retrieval effectiveness of search engines is also, as explained in the preceding chapter, influenced by the indexing processes. The reason is that words which occur frequently in the document will serve as index terms of the documents being indexed, but these words are not necessarily significant. In other words, the indexing terms might not be exhaustive and specific enough to represent the documents in the search engine databases. As a result, search engines might not retrieve the required information with maximum precision and recall, which are greatly influenced by the exhaustivity and specificity of indexing languages.

3.4.2.2. Searching and Search Engines

Search engines allow searchers to use search queries or keywords to retrieve the required information from the web. Thus, a user formulates his/her search queries based on the conceptual analysis of the information that he/she is looking for and requests or enters the search queries into the system to retrieve the information that is sought. Then the search engine retrieves the information by matching the search queries with the indexing languages of the documents in the search engine databases. In using search engines, users can formulate their own search queries by using different search strategies. In other words, users can search the required information by using Single-word search strategy, Boolean search strategy, Phrasal search strategy, Proximity, Truncation and so on.

As indexing processes influence the effectiveness of search engines, search processes also influence their effectiveness in retrieving the requested information. The reason is that most users do not come with clear information needs and thus are not be able to formulate exhaustive and specific search expressions using the various search features or search strategies. In other words, many users do not begin with such a clear view of their information requirements that they can formulate sharp search statement, while the result of a search depends heavily upon the correct understanding of the user's precise needs and the proper search queries. So, the search processes that the user follows significantly affect the retrieval effectiveness of search engines in retrieving the most relevant information from the search engine databases.

3.5. Major Search Engines

There are a number of search engines. Some of the major ones are:

- Google http://www.google.com

- Alta Vista http://www.altavista.com
- Excite http://www.excite.com
- Hot Bot http://www.hotbot.com
- Lycos http://www.lycos.com
- Wise Nut http://www.wisenut.com
- MSN Search http://www.msnsearch.com
- Teoma http://www.teoma.com
- All The Web http://www.alltheweb.com

These search engines are considered as major ones because they are well known and widely used by many searchers (Sullivan, 2003:1). They all deal with the retrieval of text documents. They also claim that they index the entire text of each web document in their databases and thus they are full text databases (Ackermann & Hartman, c1999:126). They allow keyword searching and accept the various search features, and all present their search results according to their relevance ranking.

Although most of the major search engines attempt to index the entire web, each search engine has a different way of determining which pages are most relevant to the users' search queries. The reason is that search engines have their own ways of interpreting and manipulating search expressions (Ackermann & Hartman, c1999:126). That means a relevant document may be listed second in one search engine database but it might be listed tenth in another search engine database. However, all search engines operate according to similar principles. Thus, all web pages that contain terms or words that match the users' search query will be presented in the list of results presented on screen to the user (Green, 2000:126).

In general, all these major search engines perform their task under similar conditions and principles. They all gather information from the web, index them into their databases and retrieve them later based on the searchers' request by matching the search terms with the indexed languages. Their basic objective is to retrieve the most relevant information for the users' more efficiently and effectively. But, the question here is whether all the major search engines are effective and efficient in retrieving the required information from the web. Which of them is the best one? How can the evaluation and comparison of these major search engines be done? Therefore, the next chapters deal with these questions.

3.6. Summary

Searching refers to the way the file is examined and the items in it are taken as related to a search query. Thus, it refers to the retrieval of the required information from its source database based on the similarity between users' search queries and the documents in the web databases. So, searching refers to search processes from the web that include:

- Identifying the user's information requirements
- Conceptual analysis of the contents of the information to be searched for
- Determining the "aboutness" of the contents of the information to be searched for
- Formulating search statements or queries or search languages
- Entering the search statements or queries or search terms into the system
- Reviewing the retrieved items for relevance.

Therefore, the basic task of the user in the search process is to formulate appropriate search terms. To do that, users are required to follow adequate search strategies, which are sets of decisions and actions taken during the search processes. Moreover, the user is required to adopt the various search features in order to formulate adequate search queries. Some of the features that are widely used in formulating the search statements or queries are Single-word searching, Boolean searching, and Phrase searching.

During search processes, users develop search statements that will be used to retrieve the required information. But, users lack the ability in formulating appropriate search statements or queries using appropriate search strategies and search features. In such cases, searching can influence the retrieval effectiveness of information from the web because information is retrieved based on the similarities between search terms and the items in the sources databases.

In the process of searching, users use information retrieval tools to retrieve information from the web. Some of the retrieval tools are Directories and Search Engines. Directories are predefined lists of web sites, compiled by human editors and categorised according to subjects/topics. Directories are effective in the retrieval of information from the web for general and single faceted topics because the subjects are arranged in a topical list and the user can simply follow the links to get the required information.

Search engines are computer programmes that gather information about resources on the internet by means of a robot, which is called a spider or crawler, and store this information in a database and make it accessible to Internet users through a retrieval model that allows keyword searching. Search engines are effective tools to use when users or searchers are looking for very specific information or when the search topics are multifaceted. They retrieve more and up-to-date information during the retrieval processes. However, they are influenced by the indexing and searching processes in retrieving relevant information from the web.

There are a number of search engines. Some of the major ones are Google, Alta Vista, Excite, Hot Bot, Lycos, Wise Nut, MSN Search, Teoma, and All The Web. They all deal with the retrieval of text documents. They also allow keyword searching and accept the various search features. They all present their search results according to their relevance ranking. But, each search engine has a different way of determining which pages are most relevant to the users' search queries. However, all these major search engines perform their tasks according to similar principles.

Chapter Four

Methodology for Evaluating and Comparing Search Engines

4.1. Introduction

As discussed in the preceding chapter, a number of search engines have been developed to provide access to the resources that are available on the web and users or searchers use them to retrieve the required text information from the web resources. They are all full text databases. They index and store each of the web documents on their databases. All of them retrieve the requested information from their databases by matching users' search queries with the indexed terms although each one of them has a different way of determining which search items are most relevant to the users' requests. However, the question is: how effective and efficient are those search engines in retrieving the most relevant text information from the web and which of them are more effective and efficient? So, it is very important to evaluate the performance of those search engines at retrieving relevant text information from the web. But, the main question here is: How are the evaluation and comparisons of those search engines done? Therefore, this chapter deals with the methodologies that should be followed in evaluating and comparing the performances of search engines. So, let's begin by discussing first the possible reasons for evaluating information retrieval systems.

4.2. Reasons for Evaluating IRS

There are many reasons why information retrieval systems are evaluated. Belkin & Callan (as quoted in Kowalski & Maybury, c2000:258) point out that information retrieval systems are evaluated:

- To aid in the selection of a system to procure
- To monitor and evaluate system effectiveness
- To evaluate query generation process for improvements
- To provide inputs to cost benefits analysis of an information system; or
- To determine the effects of changes made to an existing information system.

Chowdhury (c1999: 200) also states that information retrieval systems are evaluated in order to ascertain the level of their performance or their value. Thus, information retrieval systems are evaluated to see which of the existing systems perform better or to see how the level of their performance can be enhanced. In other words, they are

evaluated to see whether their intended objectives are met or not, that is, to determine the performance of systems in retrieving all the required information that are relevant to a given query while withholding non-relevant information.

However, from an academic perspective, evaluations or measurements are focused on the specific effectiveness and efficiency of a system (Kowalski & Maybury, c2000:258). So, for the purpose of this study, the evaluation was focused on the performance of information retrieval systems, in this case web search engines. Thus, the study tested the effectiveness and efficiency of the existing web search engines in retrieving relevant text information from the web.

4.3. Effectiveness and Efficiency in IRS

In an IR environment, it is very important to use some sort of scale during the evaluation study in order to measure system performances. In this case, effectiveness and efficiency are the two basic parameters used to measure the performance of information retrieval systems. Effectiveness means the level up to which the given system attains its stated objectives, whereas efficiency means how economically the system is achieving its objectives. The effectiveness may be a measure of how far it can retrieve relevant information while withholding non-relevant information, whereas efficiency can be a measure of how far the system is cost-effective, that is, at what minimum cost can it function effectively (Chowdhury, c1999: 200).

However, in order to see the efficiency of the system, it is necessary to calculate the cost factors such as response time - the time taken by the system to provide an answer, users' effort - the amount of time and effort needed by a user to interact with the system and analyse the output retrieved in order to get the correct information, the financial expenditure involved per search, and so on (Chowdhury, c1999: 200). But, in this study, only the response time was considered to test the efficiency of the web search engines because response time is a metric frequently collected to determine the efficiency of the search execution (Kowalski & Maybury, c2000:261).

4.4. Types of Search Engine Evaluation Approaches

There are two types of search engine evaluation approaches. These are called Testimonial and Shootout. Testimonials are causal studies and states the general impression obtained after executing a few queries, whereas shootouts are rigorous

studies and follow the information retrieval measures for evaluation purposes (Can, Nyray & Selvidik, 2003:3). Gordon & Pathak (1999:145-146) further explain that testimonials evaluation of search engines is based on the technical features such as the speed, ease of use, interface design or other features and their comparisons are made based on these features, whereas in shootout evaluation of search engines, different search engines are actually used to retrieve web pages and their effectiveness in doing so is compared. Therefore, in this study, the shootout type of approach was used.

4.5. Previous Studies

Salton & McGill (as quoted in Can, Nyray & Selvidik, 2003:3) states that the evaluation of text retrieval performance is a well-known research problem in the field of information retrieval. A number of individuals and groups have been evaluated the performance of the various web search engines. They also recommended some evaluation methodologies that need to be followed in the process of performance evaluation of web search engines. So, let's consider some of the previous studies on web search engines:

- Gordon & Pathak (1999) studied the performance of web search engines. They measured or evaluated the performance of eight search engines using 33 information-needs or queries. They used precision and recall parameters to measure the performance of web search engines and calculated at various document cut-off values (DCV). Then, the DCV were used for the statistical comparisons of the eight search engines. They also calculated the probability that a document retrieved by one search engine was retrieved by others as well. The findings of their study indicated that absolute retrieval effectiveness was fairly low and there were statistical differences among search engine retrievals and precision effectiveness at all document cut-values. Hence, Alta Vista and Open Text were found the best performers. But there were no statistical differences in the retrieval effectiveness among search engines for recall, although there were for precision. They also recommended some features of Web search engine evaluation as outlined in Table 4.1 from 1-7.
- Gwizdka & Chignel (1999) studied different measurements and used them to evaluate information retrievals from the web. They took the six criteria given by Cleverdon as frameworks for their study. These are: coverage, time lag, recall, precision, presentation and user efforts. However, time lag and recall were not used in the evaluation processes. To test the application of these measurements, they

evaluated three search engines with four queries. Their findings indicated that these measurements can be used to evaluate web search engines. Alta vista was found the best performer.

- Hawking, et al. (as quoted in Can, Nyray & Selvidik, 2003:3) also studied the performance of web search engines. They evaluated the effectiveness of 20 search engines using TREC-methods. They used 54 search queries. They also used precision to measure the performance of the various web search engines, but recall was not used. They calculated precision at various DCV, TREC-style average precision and mean reciprocal rank of first relevant document. Their findings indicated that there were high inter-correlations between performance measures and significant differences between performances of search engines. They also recommended more features of Web search engine evaluation in addition to the items stated in Gordon and Pathak (1999) study. See the additional features in Table 4.1 from 8-11.
- Bar-Ilan (2001) also studied the performance of search engines over a time period. He evaluated six search engines. The searches were carried out once a month for a period of ten months. His findings indicated that there is a need to study search engine stability (or rather instability) over time. Excite and Hot Bot were technically precise.
- Can, Nyray & Selvidik (2003) studied the performance of eight search engines with 25 queries. They judged the retrieved items with binary relevance judgements of users. They used precision and recall for the performance measurement at fixed DCV. Their findings indicated that a high level of statistically significant consistency exists between the automatic and human-based assessments both in terms of effectiveness and in terms of selecting the best and worst performing search engines. Hence, Alta Vista and Yahoo were found the best performers.
- Vaughan (2003) studied the performance of three commercial search engines, which are Google, Altavista and Teoma, in order to test the new proposed measurements of performance. He used four queries to evaluate the performance of the search engines. The findings indicated that the proposed measurements: quality of result ranking, ability to retrieve top ranked pages and stability comparison; can be used to evaluate search engine performances. Google performed the best.

Therefore, based on the above related studies, this study evaluated the performances of nine major search engines with four search queries. A set of measurements were

developed to measure the performance of those search engines and a statistical analysis was used to analyse the integrated data. Thus, the analysis of variance (ANOVA) and Fisher Least Significant Difference (LSD) method, which uses F-statistical test, (Montgomery, 1997:99) were used to analyse the integrated data.

Table 4.1: Desirable Features of Web Search Evaluation

1. The searches should be motivated by genuine information-needs of Web users
2. If a search intermediary is employed, the primary searcher's information-need should be captured as fully as and with as much context possible and transmitted to the intermediary
3. A sufficiently large number of searches must be conducted to obtain meaningful evaluations of search engine effectiveness
4. Most major search engines should be considered
5. The most effective combination of specific features of each search engine should be exploited (i.e. the queries submitted to the engines may be different)
6. The user who needs the information must make relevance judgments (Hawking et al. (2001) assumes that independent judges can do it)
7. Experiments should (a) prevent bias towards search engines (e.g., by blinding or randomizing search outputs), (b) use accepted information retrieval measures, (c) employ statistical tests to measure performance differences of search engines
8. The search topics should represent the range of information needs over which it is desired to draw conclusions
9. Result judging should be appropriate to the type of query submitted (e.g., some queries may need a one-line answer)
10. Document presentation should be like that of a Web browser (images should be viewable, if necessary it should be possible to follow links)
11. Dead links should count as useless answers

4.6. Features of Search Engine Evaluation

In order to make the evaluation of web search engines most accurate and informative, it is necessary to consider some desirable features. Gordon & Pathak and Hawking et al (as quoted in Can, Nyray & Selvidik, 2003:4) commented on useful search features as shown in Table 4.1. In this study, all the features given in Table 4.1 were satisfied except features 2, 3, 5 and 10. Features 2 and 10 do not apply to this study. In regard to feature 3, only small numbers of searches were conducted due to the limitation of time and human labour. Feature 5 requires expert searchers so it was very difficult to exploit the most effective combinations of specific features of search engines. Thus, the

queries that were submitted to the engines were the same. Hence, taking these features into consideration, the following search queries and relevance judgement criteria were developed.

4.7. Search Queries and Relevance Judgement

4.7.1. Search Queries

The process of measuring or evaluating retrieval performances requires search queries. So, the following four search queries, which are taken from the field of Information and Knowledge Management, were developed. Only these search queries were used to retrieve the specified information from the web and only the simple search facility of the search engines was used, not the advanced search facility.

Query 1: "Information Overload"

This is a phrase search term, which was selected to retrieve documents that discuss the concept of information overload. The retrieved documents were expected to discuss the following points:

- Definition of Information overload.
- How and when information overload occurs?
- What can be done to avoid information overload?

Query 2: "Information Systems" and "Knowledge Sharing"

This search term is a phrasal search term that is connected by a Boolean operator "AND". It was expected to retrieve documents that discuss the role of information systems in the process of knowledge sharing in an organization.

Query 3: "Knowledge Creation".

This is also a phrase search term. It was expected to retrieve documents that discuss the concept of knowledge creation in an organization. Thus, the retrieved documents were expected to discuss the following points:

- Why organizations need to create knowledge?
- How is knowledge created in an organization?

Query 4: "Deception" or "Misinformation"

This search statement was expected to retrieve documents that discuss the concept of Deception or Misinformation or both. Thus, the retrieved documents were expected to discuss the following points:

- What is meant by Deception or Misinformation
- What are their differences and similarities
- What are the advantages and disadvantages of Deception or Misinformation
- When do organizations use the concept of Deception or Misinformation

4.7.2. Relevance Judgment

In the process of searching text information from the web, the system retrieves a set of items based on the user's search queries. The user then reviews the retrieved set of items for relevance, which is the measure of the contact between a source and destination, that is, a document and its user (Chowdhury, c1999: 202) because all the retrieved items may not necessarily be significant to the users' information requirement. In this case, the user makes relevance judgments. The judgments are subjective judgments, which are made only by individuals asking the request (Foskett, c1996:12). Besides, as Kowalski & Maybury (c2000:259) states, subjective judgment depends upon a specific user's judgment and it is measurable at a point in time constrained by the particular users and their thresholds on acceptability of information. So, the evaluation of relevancy of the retrieved items is very difficult because of the users' subjectivity in the judgment. Hence, the judgment cannot be an absolute judgement.

However, for the purpose of this study, the following judgment criteria were identified in order to ensure the consistency of the relevance judgment:

- Only the first twenty retrieved items were reviewed for relevancy.
- Relevance was judged against the required information indicated in the query statements above.
- The judgment was completely independent of the other judgements (Hawking et al., 1999:1325).
- The items were considered relevant if they contained any information about the subject, irrespective of the amount of information, even if the information derived only from the context (Van der Walt, c2000:185).
- An item was not considered relevant if the search term merely occurred on the page without any useful information about the subject being communicated (Van der Walt, c2000:185).
- An item was not considered relevant if the search term merely occurred on the page as a hypertext link, unless it was obvious that the page was the home page

or main page of a resource that contain substantial information about the subject (Van der Walt, c2000:185). In other words, only the actual documents were judged. The judge did not follow links very deeply except local links, which means links within the actual documents.

- The judgment was a subjective judgment and the retrieved items were categorised as relevant/partly-relevant/non-relevant/duplicate items and dead or broken links.
- Duplicated items/links were recorded separately and they were examined or analysed separately during the analysis because the removal of duplicates from the results is something search engines should do. Duplicate items/links are defined as items/links leading to exactly the same page irrespective of whether it was found on a different server or in a different directory (Van der Walt, c2000:185).
- Pages not found were recorded separately as dead/broken links and they were examined or analysed separately during the analysis because this gives an indication of how well the search engine databases are updated.
- Since, dead/broken links are links to web pages that no longer exist, or that are inaccessible for some other reasons (Van der Walt, c2000:184), they were not counted as relevant even if it was obvious from the information on the search results page that the resource was relevant. In the context of the internet, "relevant" can be regarded as "relevant and accessible" (Van der Walt, c2000:185).
- Dead/broken links were rechecked at least twice to make sure that the inability to access them could not be attributing to server problems or other temporary conditions (Van der Walt, c2000:185).

4.8. Evaluation Criteria or Measurements

Cleverdon (as quoted in Chowdhury, c1999: 203) identified six criteria for the evaluation of information retrieval systems. These are:

- *Recall* - the ability of a system to present all the relevant items
- *Precision* - the ability of a system to present only those items that are relevant
- *Time Lag* - the average interval between the time the searcher request is made and the time the answer is provided

- *Effort* - intellectual as well as physical effort, required from the user in obtaining answers to the search requests
- *Form of presentation of the search output* - which affects the user's ability to make use of the retrieved items, and
- *Coverage of the collection* - the extent to which the system includes relevant matter.

However, the two most common measures of information retrieval effectiveness are recall, which is the percentage of the relevant items retrieved in a search, and precision, which is the percentage of the items retrieved in a search that are relevant (Sparck Jones & Willett, c1997:2; Hawking, et al., 1999:1327). Gordon & Pathak (1999:146) also states that recall measures the proportion of relevant documents in the database that are actually retrieved and precision is the proportion of retrieved documents that is relevant.

In addition to precision and recall as a measurement of effectiveness, it is also very important to examine duplicated items and broken links separately and this can be used as additional measure of system effectiveness. The reason is that retrieval systems should be able to withhold or eliminate duplicated items in the search results. In other words, the removal of duplicates from the search results is something retrieval systems should do, because they are useless to the user. In a similar manner, retrieval systems should regularly check on links and eliminate dead ones. Thus, broken/dead links give an indication of how well the retrieval system databases, in particular search engine databases, are updated.

But, precision, recall, duplicated items and broken links do not measure the efficiency of information retrieval systems. So, other measures are required to evaluate the efficiency of information retrieval systems. Therefore, as was explained earlier and as Rowley & Farrow (c2000:365-367) points out, in order to measure the efficiency of a system, it is necessary to consider the time taken to perform a search, the cost that includes any expenses associated with the acquisition of the source or access to it and the searcher's time, and usability that takes into account both interface design and the nature of indexing languages. However, in this study, only response time was used to measure the efficiency because it is a metric frequently collected to determine the efficiency of the search execution (Kowalski & Maybury, c2000:261).

Hence, precision, recall, duplicate items, broken/dead links and response time were used to measure the performance of web search engines in retrieving relevant text information from the web. They are calculated as follows:

1. **Precision**

$$\Pr ecision = \frac{Number\_\text{Re}trieved - \text{Re}levant}{Number\_Total\_\text{Re}trieved} \times 100\%$$

Where;

Number_retrieved_Relevant is the total number of relevant items retrieved by the query.

Number_Total_Retrieved is the total number of items retrieved by the query.

However, since only the first 20 retrieved items were reviewed for relevancy, the numerator and the denominator were redefined as *Number_Reviewed_Relevant* and *Number_Total_Reviewed* respectively. Hence, the formula for precision is:

$$\Pr ecision = \frac{Number\_\text{Reviewed}\_\text{Re}levant}{Number\_Total\_\text{Reviewed}} \times 100\%$$

2. **Recall**

$$\text{Re}call = \frac{Number\_\text{Re}trieved - \text{Re}levant}{Number\_Possible\_\text{Re}levant} \times 100\%$$

Where;

Number_retrieved_Relevant is the total number of relevant items retrieved by the query.

Number_Possible_Relevant is the total number of relevant items in the database.

However, it is very difficult to calculate recall because it is not possible to know the total number of relevant items in the database. However, it is possible to measure "relative recall". In relative recall calculation, the denominator term, "the total number of relevant items in the database", was replaced by "the total number of relevant items retrieved or reviewed by all search engines "(Can, Nyray & Selvidik, 2003:9). Hence, the formula for recall is:

$$\text{Re}lative\,\text{Re}call = \frac{Number\_\text{Re}viewed - \text{Re}levant}{Number\_\text{Re}viewed\_\text{Re}levant\_by\_all\_Search - Engines} \times 100\%$$

Where,

Number_Reviewed_by_all_Search-Engines are the total number of relevant items reviewed by all search engines.

3. Duplicated Items (DI) and Broken Links (BL)

A) Duplicated Items (DI)

The proportion for DI is calculated by:

$$\frac{Total\_Number\_DI}{Total\_Number\_\mathrm{Re}viewed} \times 100\%$$

Where,

Total_Number_DI is the total number of duplicated items (DI) found in the search result per query.

Total_Number_Reviewed is the total number of items reviewed per query.

B) Broken Links (BL)

The proportion for BL is calculated by:

$$\frac{Total\_Number\_BL}{Total\_Number\_\mathrm{Re}viewed} \times 100\%$$

Where,

Total_Number_BL is the total number of broken links (BL) found in the search result per query.

4. Response Time

Response Time is a metric measurement that is frequently collected in every search. Its metric units are seconds (sec) or minutes (min) or hours (hr). It was considered as the time taken to perform a search, that is, the time taken by the search engines to provide an answer. The beginning is when the user tells the system to begin searching and the end time is when the first result is available for the user to review. During the test, a stop watch was used to measure the response time.

So, based on the specified search queries, information requirements and the relevancy judgment criteria, the required experiment was conducted and the necessary data was collected and integrated. Then, the required calculation was performed and analysed

statistically. Lastly, the evaluation and comparisons were done based on the statistical results. Therefore, the next chapter deals with the evaluation and comparison results.

4.9. Summary

Evaluation and comparisons of text retrieval performances of information retrieval systems is a well-known research problem in the field of information retrieval. A number of individuals and groups have evaluated the performance of the various web search engines for various reasons. But from an academic perspective, evaluations or measurements are focused on the specific effectiveness and efficiency of information retrieval systems.

There are two types of evaluation approaches. These are testimonials and shootouts. Testimonials evaluation of search engines are based on the technical features such as the speed, ease of use, interface design or other features and their comparisons are made based on these features, whereas in shootouts evaluation of search engines, different search engines are actually used to retrieve web pages and their effectiveness and efficiency in doing so are compared. In this study, the shootouts type of approach was used to test the effectiveness and efficiency of search engines.

Parallel to the evaluation type of approach, it is very important to consider some desirable features. So, this study considered the required features in evaluating search engine performances. Moreover, the process of measuring or evaluating retrieval performances requires search queries that will be entered into the system and four search queries were used in this study. It is also identified a relevancy judgment criterion in order to make the evaluation of web search engines most accurate and informative. Additionally, it is important to use some sort of parameter in measuring or evaluating the performance of web search engines.

Effectiveness and efficiency are the two most important parameters in measuring or evaluating search engine performances. Effectiveness means the level up to which the given system attains its stated objectives, whereas efficiency means how economically the system is achieving its objectives. The effectiveness may be a measure of how far it can retrieve relevant information while withholding non-relevant information, whereas, efficiency can be a measure of how far the system is cost-effective, which means functioning effectively at minimum cost. Effectiveness can be measured by precision -

the proportion of retrieved documents that are relevant and recall - the proportion of relevant documents in the database that is actually retrieved. Efficiency of search engines is measured by the response time, which is the time taken to perform a search because response time is a metric frequently collected to determine the efficiency of the search execution. It is also very important to test the effectiveness of search engines in eliminating duplicated items and broken links because effective search engines should eliminate DI and BL. Therefore, precision, recall, duplicated items, broken links and response time were used to evaluate or measure the effectiveness and efficiency of search engines in retrieving relevant text information from the web. Hence, this study used these measurements to evaluate and compare the nine major search engines.

Chapter Five

Evaluation and Comparison Results of Search Engines

5.1. Introduction

In the preceding chapter, the methodology that one has to follow in evaluating and comparing the performance of web search engines in retrieving the most relevant text information from the web was discussed in general. Based on those methodologies, the experiment was conducted and the necessary data was collected and integrated. Then, the required calculations were computed and analysed statistically. Therefore, this chapter deals with the evaluation and comparison results of the nine major web search engines in retrieving the requested text information from the web. But, let's first discuss the experimental results.

5.2. Discussion of Experimental Results

Based on the specified search queries, information requirements and the relevancy judgment criteria, the required experiment was conducted and the necessary data was collected. Then, the collected data was integrated, as shown in Table 5.1 and Table 5.2, in such a way that the required calculations can be done. In this case, items found partly-relevant were considered as relevant and duplicated items and broken links were recorded separately. Therefore, the experiment conducted allowed me to evaluate and compare the performances of the search engines, which are their effectiveness and efficiency in retrieving the required text information from the web. So, let's see the evaluation and comparison results.

Table 5.1: Total Number of Relevant (R) and Non-Relevant (NR) Items

Search Engines	Queries								Total No. Reviewed per Query
	Query 1		Query 2		Query 3		Query 4		
	R	NR	R	NR	R	NR	R	NR	
Google	9	11	7	11	6	10	6	13	20
Alta Vista	8	12	4	15	4	15	11	9	20
Excite	7	8	4	11	4	15	4	14	20
Hot Bot	10	9	4	14	8	11	8	11	20
Lycos	8	11	5	13	1	17	3	13	20
Wise Nut	9	11	1	19	5	13	2	16	20
MSN Search	10	9	7	12	7	12	9	10	20
Teoma	9	11	4	10	4	16	1	15	20
All The Web	11	8	6	13	0	20	3	13	20
Total	81	90	42	118	39	129	47	114	720

Table 5.2: Total Number of Duplicate Items (DI) and Broken Links (BL)

Search Engines	Queries								Total No. Reviewed per Query
	Query 1		Query 2		Query 3		Query 4		
	DI	BL	DI	BL	DI	BL	DI	BL	
Google	0	0	2	0	4	0	1	0	20
Alta Vista	0	0	1	0	1	0	0	0	20
Excite	2	3	5	0	1	0	1	1	20
Hot Bot	1	0	1	1	1	0	1	0	20
Lycos	0	1	0	2	1	1	4	0	20
Wise Nut	0	0	0	0	2	0	0	2	20
MSN Search	0	1	1	0	1	0	1	0	20
Teoma	0	0	1	5	0	0	2	2	20
All The Web	0	1	1	0	0	0	4	0	20
Total	3	6	12	8	11	1	14	5	720

5.3. Retrieval Effectiveness

The retrieval effectiveness of search engines in retrieving the most relevant information from the web can be evaluated or measured by precision - the proportion of relevant documents or items judged relevant, and recall - the proportion of relevant documents or items retrieved. Therefore, based on the data in Table 5.1, the precision and recall of the search engines were computed for each query as shown in Table 5.3 & Table 5.5. So, let's discuss and analyse, first, the precision and recall results separately.

Table 5.3: Precision of Search Engines for each Query (in %)

Search Engines	Queries				Total	Average
	Query 1	Query 2	Query 3	Query 4		
MSN Search	50	35	35	45	165	41.25
Hot Bot	50	20	40	40	150	37.50
Google	45	35	30	30	140	35.00
Alta Vista	40	20	20	55	135	33.75
All The Web	55	30	0	15	100	25.00
Excite	35	20	20	20	95	23.75
Teoma	45	20	20	5	90	22.50
Lycos	40	25	5	15	85	21.25
WiseNut	45	5	25	10	85	21.25
Total	405	210	195	235	1045	

5.3.1. Precision Results

After computing the required precision of search engines for each query, the analysis of variance (ANOVA), as shown in Table 5.4, was performed based on the precision scores and accordingly $F_{Cal} \cong 2.093$ is less than $F_{Tab} \cong 2.355$ at 95% of confidence interval. Therefore, the results indicated that there were no significant differences between the search engines with regard to precision.

Table 5.4:ANOVA for Precision of Search Engines

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Search Engines	1947.222	8	243.4027778	2.092537313	0.077477342	2.35507969
Queries	3152.083	3	1050.694444	9.032835821	0.000349371	3.00878611
Error	2791.667	24	116.3194444			
Total	7890.972	35				

However, since $|F_{Cal} - F_{Tab}| = 0.262$ is too small and the average precision difference between the largest score 41.25% (for MSN Search) and the smallest score 21.25% (For Lycos and WiseNut), is relatively large enough, which is 20%, the precision of search engines was examined using the Fisher Least Significant Difference (LSD) method, which uses the F-statistics test, in order to see which group of the search engines have or do not have significant differences and to identify the better search engines in precision. Hence, according to the LSD test statistics, MSN search, Hot Bot, Google and Alta Vista were found the only ones that do not differ significantly in precision. So, they were found better in precision with MSN search (41.25%) the best one, which is then followed sequentially by Hot Bot (37.5%), Google (35%) and Alta Vista (33.75%). Lycos (21.25%) and WiseNut (21.25%) were found the least precise.

Table 5.5: Recall of Search Engines for each Query (in %)

Search Engines	Queries				Total	Average
	Query 1	Query 2	Query 3	Query 4		
MSN Search	12.35	16.67	17.95	19.15	66.12	16.5300
Hot Bot	12.35	9.52	20.51	17.02	59.4	14.8500
Google	11.11	16.67	15.38	12.77	55.93	13.9825
Alta Vista	9.87	9.52	10.26	23.4	53.05	13.2625
Excite	8.64	9.52	10.26	8.51	36.93	9.2325
All The Web	13.58	14.29	0.00	6.38	34.25	8.5625
Teoma	11.11	9.52	10.26	2.13	33.02	8.2550
Lycos	9.87	11.9	2.56	6.38	30.71	7.6775
WiseNut	11.11	2.38	12.82	4.26	30.57	7.6425
Total	99.99	99.99	100	100	399.98	

5.3.2. Recall Results

In a similar manner, the analysis of variance (ANOVA) for recall, as shown in Table 5.6, was performed based on the recall scores in Table 5.5. Just as with precision, the finding statistically indicated that there were no significant differences (since $F_{Cal} \cong 2.0613$ is less than $F_{Tab} \cong 2.3551$ at 95% of confidence interval) between the search engines with regard to recall.

Table 5.6:ANOVA for Recall of Search Engines

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Search Engines	392.8796	8	49.10995486	2.061262632	0.081651	2.35508
Queries	1.11E-05	3	3.7037E-06	1.55453E-07	1	3.008786
Error	571.8043	24	23.82518079			
Total	964.684	35				

However, since $|F_{Cal} - F_{Tab}| = 0.2938$ is too small and the average recall difference between the largest score 16.53% (MSN Search) and the smallest score 7.6425% (For WiseNut), is relatively large enough, which is 8.8875%, when compared to the computed recall results, the recall of search engines was examined using the Fisher Least Significant Difference (LSD) method as was done in case of precision. Hence, according to the LSD test statistics, MSN search, Hot Bot, Google and Alta Vista were found the only ones that do not differ significantly in the case of recall. So, once again they were found relatively better in recall with MSN search (16.53%) the best, which is then followed by Hot Bot (14.85%), Google (13.98 %) and Alta vista (13.26). Lycos (6.78%) and WiseNut (7.64%) were again found the poorest regarding recall.

5.3.3. Over All Effectiveness: Precision and Recall

In order to gauge the over all effectiveness of search engines, the average precision and recall were taken into consideration together as shown in Table 5.7. Then, the analysis of variance (ANOVA) was performed, as shown in Table 5.8, based on the average precision and recall. The finding statistically indicated that there were significant differences (since $F_{Cal} \cong 6.8223$ is greater than $F_{Tab} \cong 3.4381$ at 95% of confidence interval) between the search engines in the overall effectiveness.

Table 5.7: Average Precision and Recall of Search Engines (in %)

Search Engines	Average Precision and Recall		Total	Weighted Average
	Average Precision	Average Recall		
MSN Search	41.25	16.5300	57.7800	28.89000
Hot Bot	37.50	14.8500	52.3500	26.17500
Google	35.00	13.9825	48.9825	24.49125
Alta Vista	33.75	13.2625	47.0125	23.50625
All The Web	25.00	8.5625	33.5625	16.78125
Excite	23.75	9.2325	32.9825	16.49125
Teoma	22.50	8.2550	30.7550	15.37750
Lycos	21.25	7.6775	28.9275	14.46375
WiseNut	21.25	7.6425	28.8925	14.44625
Total	261.25	99.995	361.2450	

Table 5.8: ANOVA for the Average Precision and Recall

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Search Engines	510.2363438	8	63.77954297	6.822333694	0.00677435	3.438103136
Precision & Recall	1444.620835	1	1444.620835	154.5273757	1.6377E-06	5.317644991
Error	74.78912153	8	9.348640191			
Total	2029.6463	17				

So, the next question was: which one of the search engines caused significant differences or which one of them rated higher in effectiveness? To answer this question, the average precision and recall of search engines was examined using the Fisher Least Significant Difference (LSD) method. Hence, according to the LSD test statistics, MSN Search, Hot Bot, Google and Alta Vista are the only ones that were found non-significantly different in the overall effectiveness. Therefore, they were effective search engines and MSN Search was the most effective one followed sequentially by Hot Bot Google, and Alta Vista. Lycos and WiseNut were again rated lowest and hence they were not effective.

5.4. Effectiveness Results in Eliminating Duplicated Items and Broken Links

5.4.1. Effectiveness in Eliminating Duplicated Items (DI)

In order to see how effective search engines are in eliminating DI in the search result, the required proportion was computed for DI, as shown in Table 5.9, based on the data in Table 5.2. Then, the analysis of variance (ANOVA) for the DI was performed based on the proportion of DI results as shown in Table 5.10. Hence, the finding indicated that there were no significant differences (since $F_{Cal} \cong 0.8145$ is less than $F_{Tab} \cong 2.3551$ at 95% of confidence interval) between search engines in eliminating DI.

Table 5.9: Total number of Duplicate Items (DI) for each Query (in %)

Search engines	Queries				Total	Average
	Query 1	Query 2	Query 3	Query 4		
Alta Vista	0	5	5	0	10	2.5
WiseNut	0	0	10	0	10	2.5
MSN Search	0	5	5	5	15	3.75
Teoma	0	5	0	10	15	3.75
Hot Bot	5	5	5	5	20	5
Lycos	0	0	5	20	25	6.25
All The Web	0	5	0	20	25	6.25
Google	0	10	20	5	35	8.75
Excite	10	25	5	5	45	11.25
Total	15	60	55	70	200	

Table 5.10: ANOVA for the Duplicated Items

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Search engines	276.3889	8	34.54861	0.814461	0.597146	2.35508
Queries	194.4444	3	64.81481	1.527967	0.232813	3.008786
Error	1018.056	24	42.41898			
Total	1488.889	35				

Once again, the DI was examined using Fisher Least Significant Difference (LSD) method. According to the LSD statistics test, there were no significant differences among search engines in eliminating duplicated items. Thus, they were all effective in eliminating duplicated items. In this case, Alta Vista and WiseNut were found to be the most effective and Excite was found to be the least effective.

5.4.2. Effectiveness in Eliminating Broken Links (BL)

In order to see how search engines are effective in checking on links and eliminating the broken ones in the search result or in updating their databases, the required proportion for BL was computed, as shown in Table 5.11, based on the data in Table 5.2. Then, the analysis of variance (ANOVA) for the BL was performed based on the proportion of BL results as shown in Table 5.12. Hence, the finding indicated that there were no significant differences (since $F_{Cal} \cong 1.2049$ is less than $F_{Tab} \cong 2.3551$ at 95% of confidence interval) between search engines in updating their databases.

Table 5.11: Total Number of Broken Links (BL) for each Query (in %)

Search Engines	Queries				Total	Average
	Query 1	Query 2	Query 3	Query 4		
Google	0	0	0	0	0	0
Alta Vista	0	0	0	0	0	0
Hot Bot	0	5	0	0	5	1.25
MSN Search	5	0	0	0	5	1.25
All The web	5	0	0	0	5	1.25
WiseNut	0	0	0	10	10	2.5
Excite	15	0	0	5	20	5
Lycos	5	10	5	0	20	5
Teoma	0	25	0	10	35	8.75
Total	30	40	5	25	100	

Table 5.12: ANOVA for Broken Links

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Search engines	272.2222	8	34.02778	1.204918	0.337551	2.35508
Queries	72.22222	3	24.07407	0.852459	0.47902	3.008786
Error	677.7778	24	28.24074			
Total	1022.222	35				

However, just as with DI, the BL were examined using Fisher Least Significant Difference (LSD) method. Hence, according to the LSD test statistics, Teoma was found the only search engine that differs significantly in updating its database. Thus, Teoma was found the weakest or most ineffective in checking links and eliminating dead ones, whereas Google and Alta Vista were found most effective in checking links and eliminating dead ones. Hot Bot, MSN Search and All The Web were also found effective. Thus, they update their databases regularly.

5.5. Retrieval Efficiency

The efficiency of search engines in retrieving relevant information from the web can be measured by the response time. So, the required response time of search engines was recorded for each query as shown in Table 5.13, using a stop watch during the search query executions. Then, the required analysis of variance (ANOVA) for the response time was performed based on the recorded data as shown in Table 5.14. Hence the finding statistically indicated that there were significant differences (since $F_{Cal} \cong 13.518$ is greater than $F_{Tab} \cong 2.355$ at 95% of confidence interval) between the search engines in the response time. In other words, there were significant differences in efficiency between search engines in retrieving relevant information effectively.

Table 5.13: Response Time (in Seconds)

Search Engines	Queries				Total	Average
	Query 1	Query 2	Query 3	Query 4		
Google	1.26	1.50	1.28	1.17	5.21	1.3025
All The Web	1.87	2.05	2.14	2.27	8.33	2.0825
MSN Search	2.96	1.66	2.31	2.47	9.40	2.3500
Hot Bot	2.25	2.18	2.12	3.07	9.62	2.4050
WiseNut	2.98	5.63	1.72	3.90	14.23	3.5575
Teoma	5.46	3.57	3.41	5.10	17.54	4.3850
Lycos	4.54	5.32	4.27	3.61	17.74	4.4350
Alta Vista	2.22	8.27	2.80	6.78	20.07	5.0175
Excite	9.20	8.38	8.70	8.45	34.73	8.6825
Total	32.74	38.56	28.75	36.82	136.87	

Table 5.14: ANOVA for the Response Time

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Search Engines	157.444689	8	19.68058611	13.517505	3.383E-07	2.3550797
Queries	6.411875	3	2.137291667	1.467987316	0.2482769	3.0087861
Error	34.9424	24	1.455933333			
Total	198.798964	35				

The next question was: which one of the search engines caused such a significant difference in efficiency or which of them was most efficient? To answer this question, the response time of search engines was once again examined using the Fisher Least Significant Difference (LSD) method. Hence, according to the LSD test statistics, Google, All The Web, MSN search and Hot Bot were found the only ones that do not differ significantly in the response time. Therefore, they are the efficient search engines and Google was found the most efficient one followed sequentially by All The Web, MSN search and Hot Bot. Excite was found to be the least efficient search engine.

5.6. Effective and Efficient Search Engines

Basically, the main objective of the study was to test the effectiveness and efficiency of search engines in retrieving the most relevant text information from the web, as well as to identify which of the search engines are more effective and efficient. So, according to the statistical results discussed in the preceding paragraphs, there were significant differences between search engines in their overall effectiveness and efficiency. In other words, the finding indicated that not all search engines were effective and efficient in retrieving the most relevant text information from the web. Hence, taking all the results into consideration, it can be concluded that MSN Search, Hot Bot and Google were found the top three effective and efficient search engines. However, it is very difficult to say that they were the most effective search engines in retrieving relevant text information from the web resources. The reason is that no search engine scored or performed at a moderate level of average precision, which is a more important indicator of effectiveness (Gordon & Pathak, 1999:154), in the range of 50-60% (G. G. Chowdhury, c1999:207). However, it can be fairly said that they were the most efficient search engines because they retrieve the items from the web in seconds.

5.7. Summary

Based on the specified search queries, information requirements and the relevancy judgment criteria, the required experiment was conducted and the required data was

collected and integrated. After that, the data was analysed statistically with the analysis of variance (ANOVA) and Fisher Least Significant Difference (LSD) method, which uses the F-statistics test.

According to the statistical results, the findings for precision and recall separately indicated that even though there were not many significant differences in precision and recall, MSN search, Hot Bot, Google and Alta Vista performed better and Lycos and WiseNut were found to be the poorest. However, in the overall effectiveness, the findings indicated that there were significant differences among search engines. Hence, MSN search, Hot Bot, Google and Alta Vista were once again found better in the overall effectives and Lycos and WiseNut were found to be the poorest

In case of duplicated items, the findings indicated that most search engines were found effective in eliminating duplicated items. In this case, Alta Vista and WiseNut were found most effective and Excite was not effective enough in eliminating such duplicated items. However, in the case of broken links, the findings indicated that Teoma was the worst in checking links and eliminating dead ones, whereas Google and Alta Vista were found most effective in checking links and eliminating dead ones. Hot Bot, MSN Search and All The Web were also found effective. Thus, they update their databases regularly.

Moreover, the web search engines were evaluated and compared for their efficiency and the findings indicated that there were significant differences in their efficiency. Hence, Google, All The Web, MSN search and Hot Bot were found the most efficient whereas Excite was found to be the poorest.

In general, the findings indicated that there were significant differences in the overall effectiveness and efficiency of search engines in retrieving the most relevant text information from the web. MSN search, Hot Bot and Google were found to be the top three effective and efficient search engines, but, it is very difficult to say that they were the most effective search engines in retrieving relevant text information from the web resources although it can be fairly said that they were the most efficient search engines. No search engine scored or performed at a moderate level of average precision, which is a more important indicator of effectiveness, that is, in the range of 50-60%. Thus, the absolute retrieval effectiveness of search engines was found to be very low.

Chapter Six

Conclusions and Recommendations

6.1. Conclusions

Information retrieval deals with the retrieval of information from previously stored or recorded information. Its basic objective is to extract relevant information from the web based on the user's request. It also deals with the designing of information retrieval systems (IRS) in order to store, organize and retrieve the requested information from the web resources.

IRS are the complete organizations for obtaining, storing, and making available information to users. Their main functions are to identify, analyse and represent documents in their databases, and then retrieve the requested information by matching the user's query with document contents in the databases. So, the main aim of an IRS is to retrieve documents or information in response to a user's request in such a way that the contents of the documents or the information are relevant to the user's requirement.

Moreover, IRS are concerned with how text documents in the systems or databases are represented or indexed. The process of representing a text document with an index term for retrieval purposes is called indexing. Indexing processes must include the following specific and basic steps. These are:

- Conceptual or subject analysis of the contents of the documents being indexed
- Determining the "aboutness" of the document contents
- Identifying descriptors or keywords
- Assigning indexing terms to the contents of the document
- Organizing, recording and storing in databases for future retrieval

There are two types of indexing processes. These are called manual indexing and automatic indexing. In manual indexing, the indexing processes are done manually, where as in automatic indexing, the indexing processes are done with the help of modern computing technologies. But, both indexing processes involve indexing languages. These indexing languages are controlled indexing languages or natural indexing languages. Controlled indexing languages contain approved indexing terms in a list, and are used to assign descriptors to specific documents on the bases of

subjective interpretation of the concepts in the documents, whereas natural indexing languages are languages used for indexing documents by taking indexing terms from the document being indexed.

However, the main problem in the indexing processes, be they manual indexing or automatic indexing, is to find the content identifiers or indexing terms that can fully represent the document being indexed so that they can be matched later on with the users' search terms. This is mainly due to the lack of any logical and consistent procedures in subject analysis. In other words, there is not any mechanism to control the subject analysis of the document contents, and different indexers may analyse the contents of a given document differently because they differ in their intellectual abilities and as a result they may choose different index terms to represent the contents of given documents. In automatic indexing, the words which occur most often will serve as index terms although they might not be the most significant. So, it is very difficult to find an exhaustive and specific index language that can fully represent a document and which corresponds with the terms chosen by the user or searcher. Therefore, it can be concluded that indexing processes influence the effectiveness and efficiency of information retrieval systems, in particular web search engines, because documents are retrieved on the basis of the correspondence between the search terms expressed in a query and the index terms of documents.

In addition to indexing, information retrieval systems involve searching, which refers to the retrieval of the required information from its source databases based on the similarity between users' search queries and the documents in the web databases. Just as with indexing processes, search processes must also include the following specific and basic steps. These are:

- Identifying user's information requirement that should be searched for
- Conceptual analysis of the contents of the information to be searched for
- Determining the aboutness of the contents of the information to be searched for
- Formulating search statements or queries or search terms
- Entering the search statements or queries or search terms into the system
- Reviewing the retrieved items for relevance.

So, one of the basic tasks of the user in the search process is to formulate appropriate search terms. Therefore, users are required to follow adequate search strategies, which

are sets of decisions and actions taken during the search processes, in order to formulate appropriate search queries and view their search results. Thus, users must follow the following basic steps of search strategies. These are:

- Identify the important concepts of the search
- Choose the keywords that describe these concepts
- Determine whether there are synonymous, related terms, or other variations of the keywords that should be included
- Determine which search features may apply
- Create appropriate search statements
- Determine which retrieval systems might be used
- View the result and make a relevance judgement

Moreover, users are required to adopt the various search features or techniques in order to formulate adequate search queries, although these features have their own drawbacks. Some of the search features that are widely used in formulating the search statements or queries are Single-word searching, Boolean searching, and Phrase searching.

However, users lack ability in formulating appropriate search statements or queries using appropriate search strategies and search features. One of the reasons is that, in most cases, users do not come with clear knowledge of their information needs and as a result they do not formulate appropriate search queries. Secondly, users are not able to formulate the required search queries using the combination of the available search features. Thirdly, users lack intellectual ability in conceptual analysis of their information needs. Lastly and most importantly, users do not come with the knowledge of the indexed terms of the documents to be searched and consequently users cannot formulate search queries that exactly match with the indexed terms. Therefore, once again, it can be concluded that searching, like indexing, influences the retrieval effectiveness and efficiency of information retrieval systems in retrieving information from the web because information is retrieved based on the similarities between search terms and the items in the databases.

Nevertheless, in the process of searching or information retrieval from the web, users try to use the available information retrieval tools to retrieve information from the web for different purposes. Some of the retrieval tools are Web Directories and Search Engines.

Web Directories are predefined lists of web sites, compiled by human editors and categorised according to subjects/topics. Directories are effective in the retrieval of information from the web for general and single faceted topics because the subjects are arranged in a topical list and the user can simply follow the links to get the required information.

Search engines are computer programmes that gather information about resources on the internet by means of a robot, which is called a spider or crawler, and store this information in a database and make it accessible to Internet users through a retrieval model that allows keyword searching. They are effective tools to use when users or searchers are looking for very specific information or when the search topics are multifaceted. They also retrieve more and up-to-date information.

There are a number of search engines that have been developed to provide access to web resources. Some of the major ones are Google, Alta Vista, Excite, Hot Bot, Lycos, Wise Nut, MSN Search, Teoma, and All The Web. They all deal with the retrieval of text documents. They also allow keyword searching and accept the various search features. They all present their search results according to their relevance ranking. But, each search engine has a different way of determining which pages are most relevant to the users' search queries although they perform their tasks under similar conditions and follow the same principles. But, how effective and efficient are search engines in retrieving relevant text information from the web and which of them are more efficient and effective? An experiment was conducted to test their performances in retrieving relevant text information from the web.

Effectiveness and efficiency are the two most important parameters in measuring or evaluating search engine performance. Effectiveness means the level up to which the given system attains its stated objectives, whereas efficiency means how economically the system is achieves its objectives. The effectiveness may be a measure of how far it can retrieve relevant information while withholding non-relevant information, whereas efficiency can be a measure of how far the system is cost effective, that is, functioning effectively with minimum cost. Effectiveness can be measured by Precision - the proportion of retrieved documents that are relevant and Recall - the proportion of relevant documents in the database that are actually retrieved, whereas efficiency of search engines is measured by the Response Time, which is the time taken to perform

a search, because response time is a metric frequently collected to determine the efficiency of the search execution. It is also important to test the effectiveness of search engines in eliminating duplicated items and broken links, which are useless to the user. Therefore, precision, recall, duplicated items, broken links and response time can be used to evaluate or measure the effectiveness and efficiency of search engines in retrieving relevant text information from the web.

Furthermore, search engine evaluation requires search queries that will be entered into the system in order to retrieve the required information from the web and to test the performance of search engines. It is also required to have relevance judgement criteria in order to ensure the consistency of the relevancy judgments. Therefore, in this study, four search queries were specified with their information requirements and relevancy judgement criteria were developed. Then, based on these specified search queries, information requirements and the relevancy judgment criteria, the required experiment was conducted and the required data was collected. After that, the required proportions for precision, recall, duplicated items and broken links were computed and the required response time was recorded with a stop watch. Then, the computed proportion of the various measurements and the recorded response time were analysed statistically according to the analysis of variance (ANOVA) and Fisher Least Significant Difference (LSD) method, which uses the F-statistics test, to evaluate and compare the effectiveness and efficiency of these search engines.

Hence, according to the statistical results, the overall findings indicated that there were significant differences between search engines in their effectiveness and efficiency. In other words, not all search engines were effective and efficient in retrieving the most relevant text information from the web. Thus, MSN search, Hot Bot and Google were found the top three effective and efficient search engines, whereas Lycos and WiseNut were found to be the poorest in effectiveness and Excite was found the poorest in efficiency. However, although it can be fairly said that they were found the most efficient search engines, it is very difficult to say that they were found the most effective search engines in retrieving relevant text information from the web resources, because no search engines scored or performed at a moderate level of average precision, in the range of 50-60%, which is a more important indicator of effectiveness. In other words, the findings indicated that the absolute retrieval effectiveness of search engines was found to be very low.

6.2. Recommendations

Based on the discussions and conclusions of this study, I would like to recommend the following points:

- Since, document representation, which is indexing, greatly influences the effectiveness and efficiency of information retrieval systems, indexers should follow the basic steps of indexing processes that are outlined in the conclusion, at all times whatever the case might be, in order to ensure the effectiveness of indexing.
- Indexers should spend enough time on the subject analysis of the documents being indexed because subject analysis is what helps the indexers to determine the "aboutness" of the document and to identify the index terms. In other words, indexers should recognize that subject analysis lies at the heart of indexing processes.
- Indexers should have to consider the various types of users or searchers during the indexing processes because there are experienced or skilled users, moderate users and inexperienced or unskilled users searching information from the web.
- Indexers should make sure that the identified index terms are exhaustive and specific enough because such index terms can ensure the effectiveness and efficiency of information retrieval systems in retrieving the most relevant information. In other words, it is highly necessary to balance the level of indexing exhaustivity and specificity of the indexing languages during the indexing processes.
- Software developers should make an effort to improve the existing software in such a way that the basic indexing processes will be fulfilled adequately. Above all, the software should address subject analysis adequately and in proper ways in the process of automatic indexing. In other words, they should develop software that can ensure and perform subject analysis properly and should avoid the bias in the selection of indexing terms.
- Since, searching, like indexing, influences the effectiveness and efficiency of information retrieval systems, users or searchers are required to understand the features of information retrieval systems and follow the basic steps of search processes at all times in order to retrieve the most relevant text information.
- Users or searchers should follow the appropriate search strategies in order to ensure the retrieval of the required and relevant information.

- Users or searchers should come with clear information needs and make the required conceptual analysis in the process of searching for information from the web.
- Users or searchers should develop skills in formulating appropriate search queries using the combination of the various search features or techniques.
- Researchers should study information retrieval systems from the systems' and users' perspective simultaneously and in a coherent way because indexing and searching are the two most important factors that greatly influence the retrieval effectiveness and efficiency of any IRS.
- Researchers should develop appropriate and logical methodologies in evaluating and measuring various information retrieval systems and in particular search engines. In other words, it is very important to develop and have comprehensive and acceptable evaluation or measurement criteria and we should work very hard on that, because it can help us to develop effective and efficient search engines.
- Precision, Recall, Duplicated Items, Broken Links and Response Time can be used to evaluate and compare web search engines.
- According to the result of this study, search engine providers should be motivated to upgrade their search engine standards because the findings indicated that the absolute retrieval effectiveness was very low.
- Last but not the least, according to this study, since MSN Search, Hot Bot and Google preformed better in retrieving relevant text information from the web, I would like to recommend students and researchers to use these search engines in retrieving relevant text information from the web for academic purposes.

REFERENCES

Ackermann, E. & Hartman, K. c1999. *The Information Specialist Guide to Searching and Researching on the Internet and the World Wide Web*. Chicago/London: Fitzroy Dearborn.

Armstrong, C. J. & Large, A. (ed). 2001. *Manual of Online Search Strategies*. 3rd edition. Hampshire: Gower Publishing Limited.

Bar-Ilan,J. 2001. *Methods for Measuring Search Engine Performance over Time* [online] Available: http://www10.org/cdrom/posters/1018.pdf

Can, F., Nyray, R. & Selvidik, A. B. 2003. Automatic Performance Evaluation of Web Search Engines. *Information Processing and Management*. Article In press.

Chowdhury, G. G. c1999. *Introduction to Modern Information Retrieval*. London. Library Association

Ellis, D. c1996. *Progress and Problems in Information Retrieval*. Second edition. London: Library Association Publishing

Foskett, A. C. c1999. *The Subject Approach to Information*. Fifth edition. London: Library Association Publishing

Gordon, M. & Pathak, P. 1999. Finding Information on the World Wide Web: The Retrieval Effectiveness of Search Engines. *Information Processing and Management*. Vol 35:141-180.

Green, D. 2000. The Evolution of Web Searching: *Online Information Review*. Vol 24(2):124-137

Gwizdka, J. & Chingell, M. 1999. *Towards Information Retrieval Measures for Evaluation of Web Search Engines* [Online] Available: http://www.imedia.mie.utoronto.ca/people/jacek/pubs/webIR_eval1_99.pdf

Howaking, D. [et al]. 1999. Results and Challenges in Web Search Evaluation. *Computer Networks*. Vol 31:1321-1330.

Kowalski, G. J. & Maybury, M. T. c2000. *Information Storage and Retrieval Systems: Theory and Implementation*. 2nd edition. Boston/Dordrecht/London: Cluwer Academic Publisher.

Montogomery, D. C. 1997. *Design and Analysis of Experiments*. 4th edition. New York: John Wiley.

Rijsbergen, C. J. c1999. *Information Retrieval.* [Online] Available: http://www.dcs.gla.ac.uk/~iain/keith/index.htm or http://www.dcs.gla.ac.uk/Keith/pdf/Chapter2.pdf

Rowley, J. & Farrow, J. c2000. *Organizing Knowledge: An Introduction to Managing Access to Information*. 3rd edition. England: Gower.

Sparck Jones, K. & Willett, P. (ed). c1997. *Readings in Information Retrieval*. San Francisco, California: Morgan Kaufmann Publishers, Inc.

Strzalkwowski, T. 1995. Natural Language Information Retrieval. *Information Processing and Management*. Vol 31(3):397-417.

Sullivan, D. 2003. *The Major Search Engines and Directories* [Online] Available: http://www.searchenginewatch.com/links/article.php/2156221

Taylor, A. G. c1999. *The Organization of Information*. Englewood, Colorado: Libraries Unlimited INC.

Van der Walt, M. 2000. South African Search Engines, Directories and Portals: A Survey and Evaluation. In *Dynamism and Stability in Knowledge Organization*. Edited by Clare Beghtol, Lynne C. Howarth & Nancy J. Williamson. Vol 7. Germany: ERGON VERLAG. p. 182-188.

Vaughan, L. 2003. New Measurements for Search Engine Evaluation Proposed and Tested. *Information Processing and Management*. Article in Press.