A TECHNICAL EVALUATION OF TEN INTERNET SEARCH ENGINES FOR INDEXING AND RETRIEVING SCIENTIFIC LITERATURE

BY

BUSAGALA, L. S. P.

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF POSTGRADUATE DIPLOMA IN SCIENTIFIC COMPUTING IN THE UNIVERSITY OF DAR ES SALAAM.

DEPARTMENT OF MATHEMATICS

SEPTEMBER 2001

CERTIFICATION

The undersigned, certify that has read and hereby recommend for acceptance by the University of Dar es Salaam the dissertation titled: *A technical Evaluation of Ten Internet Search Engines for Indexing and Retrieving Scientific Literature*, in partial fulfilment of the requirements for the degree of Postgraduate Diploma in Scientific Computing.

.....

Carina K. Wangwe

(SUPERVISOR)

Date.....1/10/2001.....

DECLARATION
AND
COPYRIGHT

I, Lazaro Simon Petro Busagala, declare that this dissertation is my own original work and that it has not been presented and will not be presented to any other University for a similar or any other degree award.

Signature .....

28/09/2001

AKNOWLEDGEMENT

It was not possible to complete this work in isolation. Various people and institutions were of help. I, first, would like to express my sincere gratitude to my supervisor, Mrs C. Wangwe of Mathematics Department, University of Dar es Salaam. Her tireless, constructive criticism, help, guidance and supervision made this study a success.

I further want to thank Mathematics and Computer Science Departments Staff for the their co-operation especially, the head of Mathematics Department Dr. C. Alphonce, my lecturers, the program manager and coodinators Prof. V. G. Masanja and Dr. A. Mushi, respectively. These people dedicated all their efforts, expertise and assistance whenever I needed.

Special appreciation to my sponsor, Sokoine University of Agriculture in co-operation with the Flemish Inter University Council (SUA-VLIR project). In this regard I greatly thank the project promoter Prof. P. Niewenhuysen of the Free University of Brussels (VUB), Belgium. The Information and Communication Technologies expert Prof. E. de Smet of the University of Antwerp, Belgium. Furthermore thanks be to the project leader, the Director of Sokoine National Agricultural Library (SNAL), Mr. S. S. Mbwana. And the head of Information and Communication Technologies, Mrs D. Matovelo for their tireless and co-operative work in the process of staff development in this domain. In additional I extend my gratitude to all SUA staff without forgetting the Vice Chancellor, Prof. A. Luoga and the Deputy Vice Chancellor, Prof. P. Msolla for their readiness to give me the study leave.

I am indebted to thank Mr. D. Fuli (Systems Administrator), W. Lwiza (Computer Technician) and D. Mabula (Computing Library Assistant, SNAL) for putting smooth environment in all computer work. Also to the staff University of Dar es salaam Library (Main Campus) under the leadership of the Director, Prof. J. Nawe and associate Directors Dr. Msuya and Dr. Kiondo for providing me a smooth literature search and facilities for the experiments.

Furthermore, thanks to my fellow students and whoever was my respondent for their readiness to provide any data that was required

Last, but not least, I thank my very dedicated prospective wife Flora Yohana for carrying the burden of a very busy prospective husband at different stages of this study. Her help in typing, love and magnificent care gave me the requisite peace of mind, stamina and a fillip to work hard.

I very humbly thank my LORD, the Almighty GOD for existence, strength, and the entire blessing.

DEDICATION

To my LORD GOD and family, I dedicate this dissertation.

ABSTRACT

Several public Search Engines exist of which their coverage and response time differ. Now which one does perform best under the present connectivity? This study aimed at investigating the information indexing and retrieval effectiveness and efficiency of ten selected search engines under different connection speeds at two University Libraries namely Sokoine University of Agriculture (SNAL) and University of Dar es salaam Library (UDSM).

Google followed by Yahoo outnumbered all the other eight-search engines in terms of relevance, precision, and responsiveness. In terms of other features such as phrase searching, simple and natural language interface, high quality of display results, these search engines were the best. MetaSearch engines especially MetaCrawler performed the worst in indexing and retrieving scientific literature particularly at UDSM library. There was a significant difference of search engines performances between the two connection speeds. User information searching skills was notably poor calling an action from information professionals.

TABLE OF CONTENTS

LIST OF TABLES **Page number**

LIST OF FIGURES **Page number**

ACRONYMS

H_0 =>	Null hypothesis
H_1 =>	Alternative hypothesis
ICT =>	Information and Communication Technologies
SDRAM =>	Synchronous Dynamic Random Access Memory
SNAL = >	Sokoine National Agricultural Library (cum Sokoine University of Agricultural)
UDSM =>	University of Dar es Salaam
URL =>	Universal Resource Locator

CHAPTER ONE

1.0 Introduction

Information is an indispensable tool for each and human activity in the world. Almost all activities going on involve acquisition, processing and use of information. Universities having the role of training, research and reach-out require use of information to accomplish their endeavour.

The Internet technology has revolutionalized the way information can be acquired, processed, used and disseminated. In the literature as it will be seen in this chapter and the following other chapters, scientific literature (information) can now be accessed online in few seconds. However, it has been found that information users are overwhelmed by a lot of information most of which is irrelevant. It is an unpleasant exercise to identify and acquire information of which its availability is not known and in absent of searching tools.

To simplify the activity of searching for information, among other tools, search engines have emerged. Search Engines vary their coverage and performance. This study aimed at evaluating ten selected Search Engines in indexing and retrieving scientific literature.

This dissertation is formed of five chapters. Chapter one gives the background information to the problem. It states the problem, and provides the objectives of the study. It further presents justification of the study, states the research hypothesis, questions and the limitation of the study. Chapter two presents the literature review revealing what others have done. Chapter three elaborates the methods and materials used in the study and the data Analysis. Chapter four gives an insight of the findings and discussions on those findings. Finally chapter five includes concluding remarks and recommendations based on the findings of this study.

1.1 Background Information of the Problem

Technically the Internet is the network of many networks, all running the protocol suite connected through gateways and sharing common name and address spaces (Ruh, J F 1995; Tanenbaum, A.S 1996). It is actually millions of computers connected together in some way (phone lines, Ethernet, ISDN, cable modems) so that they share information. An increasing amount of information becomes available through the Internet for example electronic journals are replacing printed journals (Nieuwenhuysen, 1999 and Proctor, 1997). Not only electronic journals are available on the Internet but also other related information which can be utilised in various ways including academic and research purpose. De Smet and Kertens (2000) argued that Internet offers an excellent mechanism for electronic networking, with widely accessible catalogue databases, current awareness newsletters, interactive signalling and hints.

Tanzanian University Libraries such as Sokoine University of Agriculture and the University of Dar es salaam library are hooked to the Internet. The primary goal of adoption, use and application of Information and Communication Technology (ICT) in these libraries is to enable users have access to online and electronic reference materials which otherwise could not be accessed through other means. However it has been reported that users do not fully utilise the available facilities to retrieve online information. According to Augustino (2000), lack of information searching skills is the main hindrance.

Kibirige and De Palo (2000) stated that search engines provide the most common access points utilised by library/information centre users to get to electronic resources on the Internet.

Several public search engines and meta-systems exist of which their coverage and response time differ. Now which ones do perform best under the present connectivity?

This study therefore evaluated the performance of ten selected search engines in retrieving scientific information on the Internet.

1.2 Problem statement

Among the serious problems in Tanzanian libraries is the lack of adequate funds (Tweve, 2000; Wema, 2000; Mnyani, 2000). Use of the Internet in identifying and utilising the available information for academic and research seems to be cheaper as some times free articles are available online. The Internet provides convenient access to the increasing amount of literature and maximising the usage of scientific records benefits the society at large (Lawrence, 2001). He further argues that although availability varies greatly by discipline, over a million research articles are freely available on the web. If a cost is involved it is usually lower than the cost of traditional means (Proctor1997 and Levey 2001).

Although vast amounts of information are increasingly existing on the Internet, it has been reported that most users in Tanzanian Universities do not fully utilise the available facilities to retrieve the online information. One of the repeatedly mentioned reasons is inadequate information searching skills and the information on the performance of search engines on the side of users (Augustino 2000; Wema, 2000; Tweve 2000 and Katundu 1998).

The coverage of any one engine is significantly limited: no single engine indexes more than about one third of the indictable web (Lawrence and Giles 1998 and 1999). When information is needed that is present but scarce, meta-search engines tend to search for information more deeply by using several other existing search engines and collect the result before presenting to the user (Nieuwenhuysen 1999). Also the coverage of search engines differs. Often, they do return documents that do not contain the query terms. Using all search engines at a time may be clumsy and difficulty. Selecting one or more without having its performance records may be time consuming (Proctor 1997). This

study therefore investigated the information retrieval effectiveness and efficiency of selected ten search engines under different connection speeds at two University Libraries that is Sokoine University of Agriculture and University of Dar es salaam.

1.3 Objective of the Study

1.3.1 General Objective

To identify the effective and efficient search engines in indexing and retrieving documents/scientific information under two different connection speeds.

1.3.2 Specific Objectives

1.3.2.1 To determine the information users awareness of search engines and the ability to use search engines to index and retrieve scientific literature on the Internet.

1.3.2.2 To identify commonly used search engines in Tanzanian University Libraries.

1.3.2.3 To identify the most effective search engines based on the relevance of the retrieved information.

1.3.2.4 To identify search engines that support phrase query terms during searching.

1.3.2.5 To identify the most efficient search engines basing on the response time under different speeds of connectivity

1.4.0 Research Questions

1.4.1 What percentage of the information users that are aware of search engines and are therefore able to use search engines to index and retrieve scientific literature on the Internet?

1.4.2 What are the commonly (popular) search engines used in Tanzanian Public University Libraries?

1.4.3 Which search engines that are the most effective basing on the relevance of the retrieved information?

1.4.4 Which search engines are most efficient in indexing and retrieving information basing on the response time under the present connectivity?

1.5.0 Hypothesis

1.5.1 Most of the information users are not aware of search engines and are therefore unable to use search engines to index and retrieve scientific literature on the Internet

1.5.2 Users do not know how to formulate queries usable by search engines and are therefore unable to use search engines efficiently for scientific literature retrieval

1.5.3 Often, search engines do return greater numbers of documents that do not contain the query terms hence irrelevant reference materials.

1.5.4 Most of the search engines are slow, respond after waiting for a considerable long times.

1.5.5 Most search engines do not support phrase query terms or natural language of the user

1.6.0 Justification of the Study

As it has been mentioned before University Libraries face financial constraints in acquiring reference materials. Furthermore outdated information always is available in

their collections (Wema, 2000; Mnyanyi, 2000 and Katundu, 1998). The Internet technologies seem to be cheaper in acquiring information that would not be acquired through different methods (Proctor1997 and Levey 2001). Although it seems to be cheaper it is however worthless using considerable a lot of money (for instance $3,000 per month at Sokoine University of Agriculture), which is not effectively utilised as users seem to be not aware of the effective and efficient search engines under the available connection speeds.

Since there are several search engines the study will be helpful to academicians, researchers, students and other information searchers to make them appropriately select search engines to meet their information need.

The finding also will increase the literature base and information access for various purposes such as for research and other developmental programmes because evidence shows that information usage increases when access is more convenient and maximising the usage of scientific literature benefits the whole society (Lawrence 2001).

1.7.0 Limitations of the study

This study was subjected to different kinds of limitations. Among others the following were critical: -

- Time was so limited to accomplish what had already planned from the beginning. For instance it was proposed that ten queries would be formulated. Because of time limitations that is six months from design to submission it was necessary to reduce them into five queries.
- Another thing was financial constraint. There was a limited amount of funds to do the research such that the sample was forced to small.
- Respondents were some how slow to respond to my request such that 91 copies of questionnaire out of 140 were received. The main reason that was noted was that University staff and postgraduate students were very busy with University examinations.

CHAPTER TWO

2.0 Literature Review

2.1. Advantages of the Application of Information and Communication Technologies

Augustino (2000) outlined the importance of the adoption, use and Application of Information and Communication Technologies (ICT). He states that ICT enable users to have access to online and electronic reference materials. The advantages of use of ICT to access online information include the followings: -

- Ability to share resources among many users at a time. This means different users sited behind their personal computers at different locations may access the same literature materials at a time. Wide dissemination of the collected information has been enormously simplified.
- Storage of reference materials is easy. Traditionally, books, journals and other of the form was complicating and consuming a considerable large space. Currently large quantities of information can be stored in a single personal computer.
- ICT including the Internet has fastened the access methods. For instance, the information user sited in the University of Dar es Salaam library can access research papers from United States of America or anywhere in the world in few seconds.
- The information retrieval is similarly simple and fast. However access and retrieval needs the knowledge of searching skills to make it simpler.

2.2 Search Methods and Tools on the Internet

Abideen (1999) groups search methods on the Internet into mainly four types: (1) Directory search - This tools searches the information by subject matter. It is a hierarchical search that starts with a general subject heading and follows with a succession of increasingly more specific sub-heading. (2) Search engines – these search a database by using keywords or search terms or query. Search engines are programs that search for documents containing the search terms. (3) Meta-Search or multiple engine search – This kind of searching involve the utilization of a number of other search engines simultaneously. It lists the hits either by search engines or by integrating the results into a sing listing. (4) Directory with Search engines – this method of searching uses both subject and key words search methods.

Vidman (1999) claims that understanding how search tools work, selecting a search tool to use for which purposes, identifying which tool does what best offers becomes unpleasant exercise even to the most dedicated cyber searchers. According to him it involves a lot of factors to consider before doing the actual searching. Factors such as simple modes vs. advanced modes, natural language vs. Boolean logic, metasearching, portals, customization, relevancy raking, back links folders, directories and more are needed to be put into considerations by the searcher before and during searching. He urges further that the web lacks organization. Indexing and retrieving relevant information may be difficult. Emergence of search tools has been helpful in finding relevant information. Simple interface and natural language searching are features required to most of searchers.

A **search engine** is a computer program that searches for documents containing key words or phrases of interest to users. These are software programs called robots or accessible spiders that crawl through all the files on the Internet and download primarily the gathered information. Usually a search engine is comprised of two main components: A resource detection program robot, a crawler or spider is dispatched to

every site that it can identify on the web. It further downloads the pages and extracts indexed information from them to enable subsequent retrieval. The second component is a text retrieval program that access the database returning matching pages for user inspection.

2.3.1 Categories of Search Engines

Search Engines may be categorized basing on the coverage of different types of web sites and information (Sasikala and Patnak 1999 and KPL. 2001). i) General web Search Engines – These have no any limitations in terms of subjects/websites/organization. Examples of these are AltaVista and Excite. ii) Special Search Engines covering specialized subject information sources and web sites of specialists organization and groups like GovBot (Covering government web sites), DejaNews (Covering different Newsgroups). ii) Metasearch engines which allow the user to access multiple search engines from different databases such as MetaCrawler and SavvySearch. This category does not own database and present results using a single interface. Although (i) and (ii) categories are separated, they can be grouped into a single search engine.

2.3.2 Architecture of search engines

2.3.2.1 Standard Search Engine

Standard search engines often poorly approximate the information need leading to low coverage by using a query. According to Lawrence (2001), the query is applied to a local database of Web pages and results are ordered and shown to the user. Most search engines have a single ordering policy i.e. all users with the same query get the same results presented in the same order. Figure 1 diagrammatically represents standard search engine architecture.

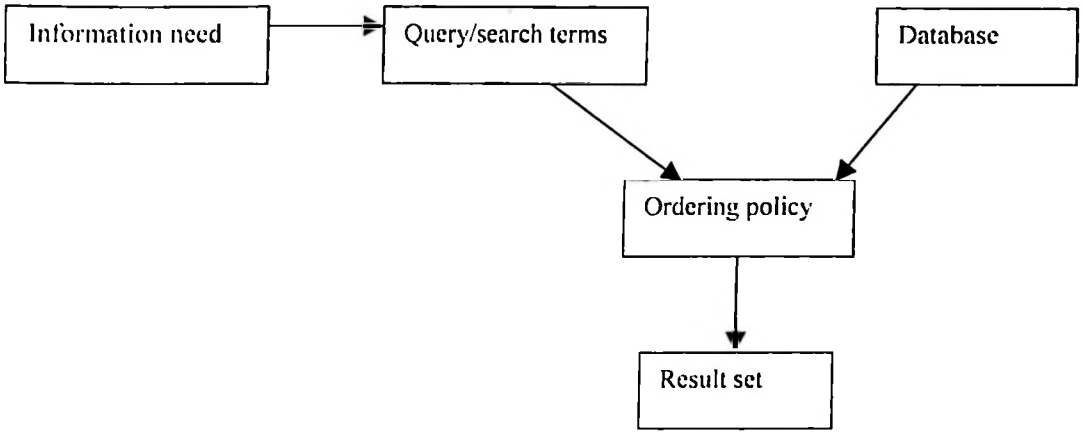

Figure 1 Standard search engine (Source: Lawrence 2001)

Abideen (1999) presented some problems with the usage of Single (standard) search engine (e.g. AltaVista, Infoseek): Difference in search syntax – Every Search Engine may have particular search syntax; Incompleteness of the database of web sites. Single Search Engines do not cover the whole World Wide Web. They may leave out relevant documents required by the user, difference in frequency of updating and search capabilities and the difference in the display and search interface.

2.3.2.2 MetaSearch Engines

Unlike a standard search engine a Metasearch engine does not have a local database and relies on other sources (i.e. on other search engines to increase information search coverage), as shown in the figure 2 below.

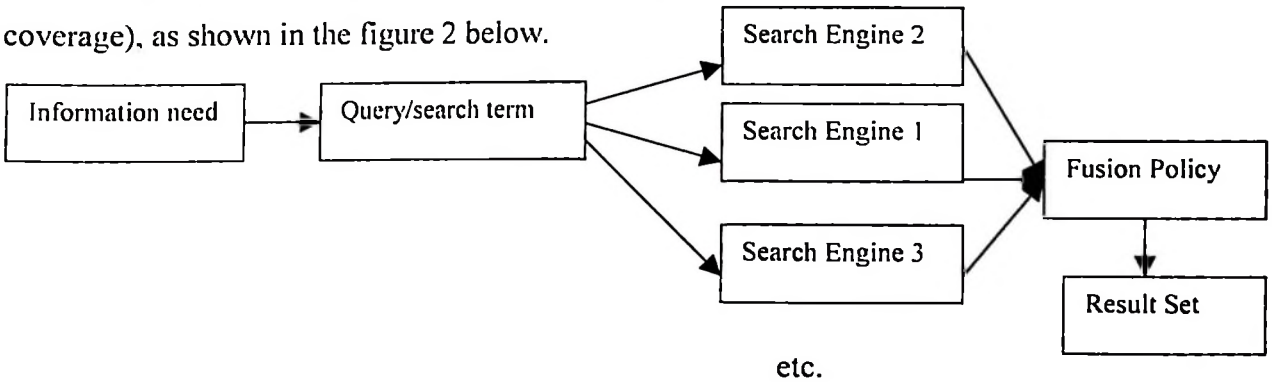

Figure 2 Architecture of a standard Meta Search engines (Source: Lawrence 2001)

The results brought from the other sources, using some other search engines, are combined through some combination policy, which is also called a fusion policy. Meta search engines typically consider only the titles, summaries, and URLs provided by the sources.

Advantages of using metasearch engines include: (i) use of only one interface for searching is to be learned (ii) Only one search query is needed to search in many engines. (iii) More thorough searching access a wider numbers of search engines (iv) An integrated set of results, in many cases, without duplication. Limitations of metasearch engines include: (i) They are subject to timeout, when search processing takes too long. (ii) Total number of retrieved information may be less than that found by engine. (ii) Most of them do not cover all popular search engines. (iii) Phrase and Boolean searching may not be properly processed or available. (iv) Advanced search facilities of individual search engines are usually not available (Abideen 1999).

2.3.2.3 Public search engines and Commercial Search Engines

Bar-Ilan (2000) did a content analysis. He analyzed the Web pages retrieved by the major search engines on a particular data as a result of the query 'informetrics OR informetric'. The list of references obtained form the Web was compared to data retrieved from commercial databases. For most cases, the list of references extracted from the Web outperformed the commercial, bibliographic databases. The results of these comparisons indicate that valuable, freely available data is hidden in the Web waiting to be extracted from the millions of Web pages.

2.4.0 Searching and Evaluating Retrieved Documents on the Internet

Explosion of available information on the Internet has made the Web search time consuming and complex process (Lesser 2000; Lawrence, 2001). Most people are

experiencing information overload. The ease with which electronic information can be stored, disseminated and manipulated threatens rather than alleviating information overload. This situation asserts that there are must be methodologies to enable people to access relevant information at time when it is needed (Hanka and Fuka 2000; Proctor, 1997). Sasikala and Patnaik, (1999) argued that the Internet has abundance of data making information retrieval a challenge on how relevant data can be retrieved precisely in a short time.

Kibirige and De Palo (2000) reported that the coverage of search engines is not about a half and results may be misleading due to the following reasons: - (1) Documents may have been changed after they have been picked up for inclusion. (2) Deleted materials may be displayed as available. (3) Web-sites that are passwords accessible are not covered. (4) The formation retrieval technology used by search engines that may not require exact match of the query content. For example Excite uses concept-based clustering and Infoseek uses morphology concept such that they may return documents with related words. Introna and Nissenbaum (2000) urged that research results suggest that search engines systematically exclude, in some cases by design and some accidentally certain web sites and index some web sites in favour of others, systematically giving prominence to some at expense of others. Exclusion differs from one search engine to another leading to the need of evaluating their performance and to present the results to the information professionals and users.

There have been various ways used to evaluate relevance of the documents retrieved from information systems. Allen (2001) adopted some of them. The number of retrieved documents that were read for a relatively long time by the users were implicitly considered relevant and those, which were printed and acknowledged by the information users, were explicitly considered relevant. According to him relevance can also be

evaluated through counting only the terms used in searches that are viewed more than once. This can be considered as technical relevance.

The user posing the query may judge the relevance of the document. This is the kind of judgement that may be considered as a subjective measure. In the context of the World Wide Web, usually only the first ten of twenty documents are examined for relevance and computed for precision. This is due to the fact that web users hardly ever go beyond the first twenty hints and that, search engines rank their results according to relevance. The documents that have all and more words of the search terms are ranked highly (Bar-Ilan 2000; Lawrence, 1998). Technical relevance is measured by considering the retrieved document relevant if it contains all search terms or phrases. Furthermore the such a document should miss all terms or phrases that are supposed to be missing that is terms preceded by minus sign or a NOT operator. Technical relevance is objective and easily checkable measure (Bar-Ilan 2000).

Precision is another measure of information retrieval systems (Salton, 1989). It is a percentage of relevant retrieved documents out of the total number of documents retrieved by the system on a query.

End users searchers may search infrequently and usually have fewer information skills than their professional information colleagues do. A new range of databases and search engines with natural language interfaces are required to assist end users (Weimer and Rusch 1996; Tomaiuolo and Packer 1996). Simple interfaces and natural language searching that respond to the needs of the typical searcher are required (Vidmar 1999).

CHAPTER THREE

3.0.0 Methodology

Two methods were adopted namely experimental tests and survey.

3.1.0 Sources of Data

Two kinds of data were involved in this study that is secondary data and primary data.

3.1.1 Primary Data

Sokoine University of Agriculture Library and University of Dar es salaam were chosen as places to do the experiment.

Sokoine University of Agriculture Library (Sokoine National Agricultural Library (SNAL)

As the name indicates this library is located at Sokoine University of Agriculture in Morogoro. Tanzania. It is also a National Agricultural Library since 1991 under the act number 21 of the same year. It provides its service to the University community and to the public at large in the field of Agriculture and other related subjects. This Library is connected to the Internet and the computerisation process is still at infancy stage. It shares its connection speed with the entire University, which also shares 1Mbps with several other universities. Currently it is about to expand its services as far as Literature Internet search is concerned. So it was appropriate to evaluate search engines so that recommendations are to be presented to the information professional and the normal information users.

University of Dar es Salaam Main Library

This Library is located at the main campus of the University of Dar es Salaam. It mainly offers its service to the University community and the public at large. It is also hooked to the Internet. It shares its 1Mbps connection speed with other campuses that include Mhimbili College of health sciences. University College of lands and architectural sciences and Hubert Kariuki University. Internet services are offered to the users in reference department. The use and application Information and Communication Technologies is not yet advanced meaning that computerisation process is in progress.

3.1.2 Secondary Data

Secondary data was obtained from the University of Dar es Salaam and Sokoine University of Agriculture Library. Local collections and online information was included. Review of documents and relevant literature was done.

3.2.1 Constraints put to the experimental tests

In order to get high quality findings the following were put in considerations: -

i) Random errors – errors that are irregular with respect to time and hence can not be predicted. These were minimized by recording five values of the same query (index and retrieval). Finally taking the average (mean) of all values (Chatfield 1983).

ii) Counting search terms used in searches that were viewed in retrieved documents.

iii) Five (5) queries were run on ten search engines and number of results returned by each engine was reported. This approach was important to minimise the error of making wrong inferences

iv) In query formulation more than one or two words is desirable for more relevant

and precise results. One word query with very common item ends up with millions of results of which most of them are irrelevant (Becket 1998). Therefore a query should have more extra words.

v) It was important to know the available bandwidth at each time during the experimental tests. This was important because search engines may be limited by the available network bandwidth (Lawrence 1998). For every one and half-hours the available bandwidth was examined by using the online services provided by search.com. The average bandwidth was 25.5 and 9.58 kbps at Sokoine National Agricultural Library and University of Dar es Salaam Library respectively.

vi) The browser used was Netscape mainly version 4.5 to avoid errors that could be caused by the use of different browsers.

3.2.2 Specification of Computers used in the experiments

The computer used at Sokoine National Agricultural Library had the following specifications: -

- Processor speed 366Mhtz
- SDRAM 32MB
- Storage Disk Space 4.3GB

The computer used at the University of Dar es Salaam had the following specifications: -

- Processor Speed 366Mhtz
- SDRAM 64MB
- Storage Disk Space 4.0GB

3.2.3 Selection of search engines to evaluate

It was important to select ten or less search engines due to the fact that time wouldn't be enough to work with more search engines (The selected ten search engines are presented

in appendix C). Most researchers who probably worked with the search engines did their research at a different connection speed. Consideration was taken on the fact that the University libraries dealt with, have financial constraint. Users always choose search tools that are accessible at no cost. Due to this fact it was reasonable to evaluate search engines that are available to users at free of charge. Chu and Rosenthal (1996) argued that these free services might continue to be available to the Internet community in the foreseeable future.

3.2.4 Query Formulation

Five queries were selected basing on the information needs of five postgraduate students of Sokoine University of Agriculture. The queries consisted of the following words: -

Query 1-> tephrosia vogelii maize growth yield

Query 2-> gypsum bean yield salt soils

Query 3-> wood ash rice straw nutrient ruminant

Query 4 -> azolla caroliniana paddy gield

Query 5 -> ethnoveterinary aloe

Since every search engines may require a specific syntax in order to retrieve relevant results consideration was taken in this regard.

3.3 Data Collection for the Experimental Tests

The experimental tests were done between 14 and 30 in July 2001. The formulated queries were adopted. The syntax of the queries slightly varied depending on the recommendations of the search engine developer in their documentation's. However phrase query experimentation was done to each

search engines. For each query five tests were done. Counting of indexed and retrieved documents was done to obtain the total number of documents.

Consideration of relevant documents was done up to the twentieth document because in context of the World Wide Web. usually on the first ten or twenty documents are examined for relevance and computed for precision. This is due to the fact that web users hardly ever go beyond the first twenty hints and that. search engines rank their results according to relevance. The documents that have all and more words of the search terms are ranked highly (Bar-Ilan 2000: Chu and Rosenthal, 1996; Dong and Su. 1997: Salton. 1989)

3.4 Data Analysis for Experimental Tests.

Data from experimental tests were summarized and computations of various quantitative measures were done. Values such as mean precision mean response time and number of retrieved relevant documents were obtained. Drawing of various figures and table was done to present the results in more desirable format. Furthermore statistical tests for significance of results were adopted. In the statistical tests t-distribution was adopted because of the following conditions: the sample size n was less than 25 or 30. The standard deviation σ was not known and the assumption that the population from which the sample is drawn was approximately normally distributed (Mann 1995, Chatfield 1983).

3.5.0 The Survey

3.5.1 Area of study

Sokoine University of Agriculture and the University of Dar es Salaam was chosen to be the area of study. These are all public universities. They are well established compared to other Universities in Tanzania with well-qualified academician and researchers. Their locations are described in section 3.1.1.

3.5.2 The population

This included university academic or research staff and postgraduate students. The assumption made was that these groups of users are highly involved in literature search to fulfil academic and research purposes.

3.5.3 Sampling and Sample Size

The sampling procedure involved random and purposive sampling. The later was employed in deciding that respondents had to come from a variety of departments or institutes. The former was applied in randomly choosing a respondent in any chosen department. Finally 140 respondents were sampled.

3.5.4 Data Collection

A structured questionnaire was used to collect the data (Appendix B). 140 copies of the questionnaire were distributed. Among them 70 copies were given to the respondents at Sokoine University of Agriculture and the remaining at the University of Dar es Salaam. The response rate was 65%, which is equivalent to 91 copies of questionnaires.

3.5.5 Data Processing and Analysis

The data was coded and entered into the computer using SPSS program, which was used to analyse the data. Frequencies and cross tabulation was done to obtain the summaries and information usable to present the results.

CHAPTER FOUR

4.0 Results and Discussion

This chapter presents the findings from the study. As far as the study is concerned it was conducted using two methods namely experiments and a survey that sampled information users from the two Universities in question. Therefore results and discussion are given in two categories basing on the experiment and the survey.

4.1.0 Results from the Experimental Tests

4.1.1 Relevancy or Precision Evaluation

Under this study it was aimed at evaluating the performance of search engines for indexing and retrieving scientific literature. Table 1 shows the precision of each evaluated search engine in the two connection speeds. Google had the highest precision at both SNAL and UDSM Library i.e. 95% and 94% respectively. Yahoo was the second with 94% at SNAL and 88% at UDSM library. Generally Webcrawler had relatively low precision at both places. The trend of retrieval precision is presented in figure 3.

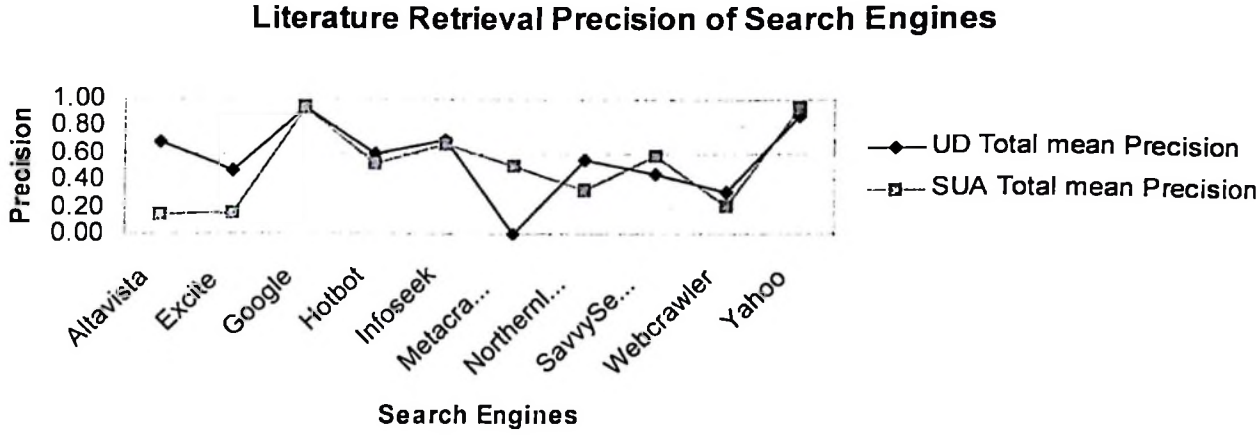

Figure 3

MetaCrawler was found to retrieve and index nothing at UDSM library, it was subject to timeout for every test carried out. This finding was supported by the argument made by Abdeen (1999). This situation may often happen to environments of slow speed like that of UDSM library.

Table 1: Precision of Search Engines at Sokoine National Agricultural Library (SNAL) and University of Dar es Salaam Library (UDSM).

Serial No	Search Engine	UDSM Precision	Mean SNAL Precision
1	Altavista	0.69	0.13
2	Excite	0.48	0.16
3	Google	0.94	0.95
4	Hotbot	0.60	0.53
5	Infoseek	0.70	0.67
6	MetaCrawler	0.00	0.51
7	Northernlight	0.55	0.33
8	SavvySearch	0.45	0.59
9	WebCrawler	0.31	0.21
10	Yahoo	0.88	0.94

Source: Study 2001

4.1.2 Responsiveness of Search Engines

Responsiveness refers to how quickly do the search engines respond from when the query is submitted to the point of displaying the results as well as the time in which the interface loads the hits. Table 2 shows the mean response time at each connection speed. Google had the shortest response mean time at both connections that are 8.09 seconds at University of Dar es Salaam library and 10.64 seconds at Sokoine National Agricultural Library. MetaCrawler had the longest response mean time at University of Dar es Salaam because the connectivity was relatively slow making it subject to time out.

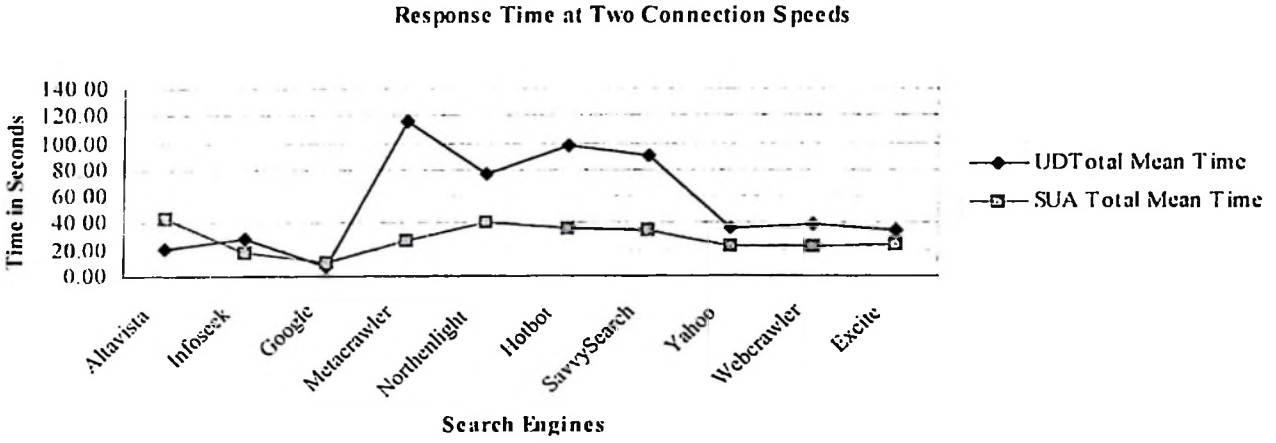

Figure 4

The trend at all connection speed is graphed in figure 4 above. It can generally be observed that most of the search engines responded in more short time (quickly) at Sokoine National Agricultural Library than at University of Dar es Salaam.

Table 2: Response Mean Time of each Search Engine at Sokoine National Agricultural Library and University of Dar es Salaam Library.

Serial No	Search Engine	UDSM Mean Time in Seconds	SNAL Mean Time in Seconds
1	Altavista	21.49	43.61
2	Excite	34.12	24.25
3	Google	08.09	10.64
4	Hotbot	97.19	35.64
5	Infoseek	28.69	18.41
6	Metacrawler	116.20	26.58
7	Northernlight	77.12	41.14
8	SavvySearch	90.08	34.10
9	Webcrawler	39.24	23.21
10	Yahoo	36.42	22.77

Source: Study, 2001

4.1.3 Inferences of Samples under Experimental Tests

4.1.3.1 Responsiveness of Search Engines

After noticing differences in terms of time response at the two connection speeds, making inferences to the whole situation seemed to be essential. The null hypothesis (H_0) assumed that there was no difference in terms of response mean time at Sokoine National Agricultural Library and those at the University of Dar es Salaam. Conversely, the alternative hypothesis (H_1) assumed to have a difference. The test statistics at 5% significance level revealed that generally there was significant difference in terms of time response of search engines at the two connection speeds. The two-tailed test $P(|t_0| = 2.21)$ was larger than probability of t-distribution $P(|t| > 2.10)$. With this results it is generally taken to be reasonable evidence that the null hypothesis (H_0) was untrue. The implication is that there was a significant difference in terms of responsiveness of search engines at the two connection speeds. In other

words search engines responded at significantly higher speeds at the Sokoine National Agricultural Library than at the University of Dar es Salaam Library.

4.1.3.2 Significance Tests on Precision

Significant tests on precision were done. The null hypothesis (H_0) claimed that there was no difference in terms of precision of search engines at the two connections speeds namely Sokoine National Agricultural Library and University of Dar es Salaam Library. The alternative hypothesis (H_1) in return claimed that there was relatively high precision at Sokoine than at University of Dar es Salaam Library. It was found that at one-tailed test $P(t_o = -18.97)$ was larger than probability t-distribution $P(t_{0.01}, 18 = 2.552)$. This means the results were significantly different at 1% significance level. In other words generally search engines were significantly more precise at the Sokoine National Agricultural Library than at University of Dar es Salaam Library.

4.1.3.3 Significance Tests for Meta Search Engines

MetaCrawler and SavvySearch were the Metasearch engines selected for evaluation.

MetaCrawler

Logically, it was unnecessary to test the significant differences at the two connection speeds. The main reason was that MetaCrawler at the University of Dar es Salaam was completely unable to retrieve any document (see figure 3). MetaCrawler had also the longest periods to respond such that it was subjected to time out (see figure 4).

SavvySearch

Significant tests on number of retrieved documents were conducted. The null hypothesis (H_0) claimed that there was no difference in terms of number of documents retrieved by SavvySearch at Sokoine National Agricultural Library and those at University of Dar es Salaam Library. The alternative hypothesis (H_1) stated that there was a difference in terms of number of documents retrieved at Sokoine National Agricultural Library and those at University of Dar es Salaam Library.

It was found that at two-tailed test statistics $P(|t_o| = 2.78)$ was larger than probability of t-distribution $P(t_{0.025}, 8 = 2.306)$. The argument made is that, the number of relevant documents retrieved at Sokoine National Agricultural Library was significantly higher at 5% significance level. This is to say SavvySearch significantly retrieved larger numbers of relevant documents at Sokoine National Agricultural Library than at University of Dar es Salaam Library.

Responsiveness Significance Tests

SavvySearch responded quicker at Sokoine National Agricultural Library than at University of Dar es Salaam that is 34 seconds and 90 seconds respectively. This figure indicate that SavvySearch was capable of responding more quickly in relatively higher speeds (see figure 4 and table 2).

4.1.4 Evaluation of Other Features

Features of ten selected Search Engines were evaluated (see Table 3). The features included the capabilities for phrase search, Boolean search, ability to restrict search to a particular time frame and quality displayed search results. Here are the results and discussion for each search engines.

Table: 3 Some features and Searching Mechanism

	Advanced Capabilities			
Search Engine	**Phrase Search**	**Boolean Search**	**Restrict to time frame**	**Display of Results**
AltaVista	Poor, require Adding "+", "-"	Well Supported	Supported	Summary of Document with Attribution
Excite	Supported	In-built the user	No	Mostly are web sites
Google	Very well supported	In-built flexible well perform	Yes	Excellent summary search terms in bold.
Hotbot	Supported	Well supported	Yes, ten years	Summary with Attribution
Infoseek	Well supported	Supported	No	Summary with Attribution
MetaCrawler	Supported	No	No	Summary with Attribution
Northernlight	Supported	Supported	Yes	Summary with Attribution
SavvySearch	Supported	Poorly supported	No	Summary with Attribution
WebCrawler	Poorly supported	Well supported	No	Web sites or Summary
Yahoo	Well supported	Well supported	Yes, Not more than four years back	Good, summary with attribution search terms appear in bold

Source: Study 2001

AltaVista produced poor results on queries that had no Boolean operators that is it could bring false hits that seemed to be highly ranked. Phrase searching for desired results would be improved by use of Boolean search. A query containing a group of words without double quotes or other Boolean operators could index millions of pages most of which were irrelevant. In general AltaVista do not support natural language queries. The ability to restrict search to time frame was moderately supported. The results display contained abstraction with attribution. However, search terms were not obviously seen in bolds making the technical relevance judgement difficult.

Excite supported phrase searching with some difficulties in such a way that it required use of Boolean operators for precise results. In the Advanced search option Boolean searching is in-built. The user is required to select whether it should use "AND", "OR", "NOT" and other operators such as quotation marks, plus (+) and minus (-) signs. Restriction of search to time frame is poorly or not supported. Results are mainly in terms of web sites. Although results may show titles or URL, they are mainly links and web sites. Search terms are not obviously displayed in results.

Google highly supported phrase searching without a requirement of the use of Boolean operators. By using more words in a phrase query it highly improved the relevance of the results. It can be said that using Google in searching you will rarely require many techniques in query formulation. This follows the fact that it ranked highly relevant documents by using normal words without demanding a special syntax. Advanced search option offered an interface, which seemed to have in-built Boolean operators, option could be chosen using radio buttons. The user was required to select the options such as; Find results with ALL of the words, the EXACT PHRASE, with ANY of the words and WITHOUT the words. It showed ability to restrict search to web pages or documents in other formats that had been updated in one year ago. Another feature was to find results updated at any time. Results displays by Google were in an excellent format. Results contained a summary with attribution. The term excellent format in this context implies that search terms (words) appeared in bold simplifying

the process for technical relevancy evaluation. It was possible to explicitly tell which document was technically relevant. Further more results display contained the format indication. It was easy to identify whether the document was in HTML or PDF formats.

Hotbot moderately supported phrase searching in terms of producing highly ranked relevant documents. Boolean search was well supported such that it increased the number of relevant documents and decreased irrelevant documents. Hotbot showed high capability in restricting search to time frame. It was possible to restrict search to index documents present on the Internet for about ten years ago. The results display was generally not excellent

Infoseek well supported phrase queries. However, it brought a limited number of hits that is between 0 to 15 hits. Boolean search was moderately supported because it rarely improved the results. The main weakness with this it couldn't show any ability to restrict search to time frame. The results display contained title summary with attribution. Search terms were not in bold creating difficulties in direct evaluation for relevancy.

MetaCrawler was expected to bring more desirable results because it is highly recommended in the literature. Not only is it highly recommended but also MetaCrawler is a Meta-Search engine that sends searches to many search engines and display results on single interface.

Conversely, MetaCrawler had a moderate performance, it moderately supported phrase Boolean search. Restriction to time frame was not supported. Abstraction of result display was found with attribution. Search terms were not bolded hence difficulties in relevance evaluation.

Northernlight supported phrase and Boolean search. It was capable to restrict search to time frame. Results display consisted of title, summary with attribution. However, search terms were not all shown in bold creating a problem in the process of evaluation relevance.

SavvySearch was among the MetaSearch engine evaluated. It also revealed moderate performance. It supported phrase search. Boolean search was poorly supported. It did not show ability to restrict search to time frame. Search results were displayed with the title link, abstract of the document found with attribution. A common problem was that the search terms were not obviously indicated in bolds.

WebCrawler supported phrase searching. However, high numbers of irrelevant materials were indexed. Boolean search could reduce the number of irrelevant materials. Restriction of search to a specific time frame was not evidently noticed. The hits produced were mainly leading a web site not the particular document as in other search engine. This situation was also evident with Excite.

Yahoo well supported both phrase and Boolean search. It manifested ability to restrict search to a specific time frame -not more that four years ago. Results display was generally good consisting of title, abstract with attribution. An interesting quality of result display was that search terms were in bold reducing the difficulties of evaluating the relevance of indexed materials.

4.2.0 Results from the Survey

4.2.1 Use of Internet in Literature Search

Table 4 below indicates that most of the academic or research staff and postgraduate students (92.3%) use the Internet in Literature Search. This shows that Internet has become an indispensable service to the university communities. But how effectively and efficiently do users use the Internet? The question may remain unanswered. The rate of using the Internet is also presented in the same table. 27.5% of the respondents do not use the Internet frequently. 51.6% use the Internet frequently and 13.6% use this service very frequently.

Table 4: **Use of Internet and Search Engine in Literature Search**

Variable		Frequency	Percentage
Use of Internet in Literature Search		84	92.3
Rate of use	Not frequent	25	27.5
	Frequent	47	51.6
	Very frequent	12	13.6
	Don't use	07	07.7
Heard of Search engines	Yes	66	72.5
Knowledge of Query Formulation	Know	26	27.5
	Don't know	66	72.5
Know Search Engine role		59	64.8

Source: Study 2001

4.2.2 Awareness about Search Engines and Query Formulation

Although it was revealed that most of the respondents use the Internet in literature search, relatively low percent (72.5%) of them had ever heard of search engines (see table 4 above). This is a challenge to the information professionals. This situation may call for review of methods used to enrich the university communities with information skills. 72.5% had implicitly no any query formulation skills. The implication is that there is a high possibility that information users spend a lot of time searching for literature with limited relevant results.

4.2.3 Commonly Used Search Engines

With the use and application of information and Communication Technologies, it was interesting to identify the mostly used search engine. Table 5 below indicates the frequency and percentage distribution for each search engines. It is important to note that only the respondents' frequencies that use the particular Search Engines are presented. The assumption made is that the left out percent of respondent do not use the listed search engines. Further more it is good to note that the search engines that do not appear in the list are not used or are rarely used by very few people.

Yahoo was found to be the most popular as it had 67% of the respondents. Google was the second; it had 49.5% of the respondents. AltaVista ranked in the highest three with 44% percent of the respondents.

Table 5: **Distribution percentages of the Search Engines used in Areas under Study**

Search Engines	Frequencies	Percentage of Respondents
AltaVista	40	44.0
Excite	17	18.7
Britannica	01	01.1
Cyber	01	01.1
Google	45	49.5
Hotbot	09	09.9
Infoseek	14	15.4
Lycos	16	17.6
Meta Crawler	04	04.4
Open Text	03	03.3
Searchalot	05	05.5
Snap	01	01.1
WebCrawler	04	04.4
Yahoo	61	67.0

Source: Study 2001

4.2.4 Preferred Search Engines

As far as preference is concerned 37.4% of the respondent prefer most Yahoo. This was the highest percent.

Table 6: **Preferred Search Engines**

Search Engines	Percentage of Respondents
Yahoo	37.4
Google	20.9
AltaVista	08.8
Others	01.1-02.2

Source: Study 2001

Google was the second with 20.9%. AltaVista being the third had 8.8% of respondents. The rest of search engines had a range of 1.1% - 2.2% respondents. 16 respondents did not respond to this question.

4.2.5 Respondent's Literature Search Performance

It was also important to know whether the respondents get the information they need or not. This parameter was included to implicitly correlate with their searching skills especially the use of search engines and kind of queries they formulate. Table 6 below shows Respondents' Performance of Literature Search. 51 respondents (56%) are overwhelmed with irrelevant materials whenever they search for literature. The implication of this figure may be that a good number of information users do not know or have limited searching skills. The evidence is presented in table 4 where it was found that 72.5% of the respondents acknowledged that they don't know to formulate queries, only 27.5% knew phrase and possibly Boolean query formulation.

Table 7: **Literature Search Performance**

Performance	Frequency	Percent
Get literature with a lot of irrelevancy	51	56
Get literature with limited results	11	12.1
Get literature without difficulties	29	31.9
Don't get any literature	1	1.1

Source: Study 2001

A critical evaluation reveals that although the connection speeds are generally slow the Internet can offer relevant reference materials just as it was stated in literature review (in chapter two). The table 7 above indicated that only 1 respondent said to get no

literature in his/her search. which is probably due to lack of information searching skills.

CHAPTER FIVE

5.0 Conclusion and Recommendations

5.1.0 Conclusion

Google outnumbered all search engines scoring the highest precision at both SNAL and UDSM Library i.e. 95% and 94% respectively. Yahoo was the second with 94% at SNAL and 88% at UDSM library. Meta Crawler was found to retrieve and index nothing at UDSM library, it was subject to timeout for every test carried out. This finding was supported by the argument made by Abdeen (1999). This situation happens to environments of slow speed like that of UDSM library. The hypothesis that most of search engines return a greater numbers of irrelevant materials or documents was supported in this regard.

Google had the shortest response mean time at both connections that are 8.09 seconds at University of Dar es Salaam library and 10.64 seconds at Sokoine National Agricultural Library. MetaCrawler had the longest response mean time at University of Dar es Salaam because the network performance speed was relatively slow making it subject to time out. This means, metasearch engines like metacrawler seem to perform poorly under slow connection speeds. The results were significantly different at 1% significance level. In other words generally search engines were significantly more precise at the Sokoine National Agricultural Library than at University of Dar es Salaam Library. Under this area of study the hypothesis that most of the search engines are slow was true. The trend at all connection speeds is graphed in figure 4 above.

It can generally be observed that most of the search engines responded in more short time (quickly) at Sokoine National Agricultural Library than at University of Dar es Salaam. The test statistics at 5% significance level revealed that generally there was significant difference in terms of time response of search engines at the two connection speeds. The two-tailed test $P(|t_0| = 2.21)$ was larger

than probability of t-distribution $P(|t| > 2.10)$. With this results it is generally taken to be reasonable evidence that the null hypothesis (H_o) was untrue. The implication is that there was a significant difference in terms of responsiveness of Search Engines at the two connection speeds. In other words Search Engines responded at significantly higher speeds at the Sokoine National Agricultural Library than at the University of Dar es Salaam Library.

Google supported exceedingly well the phrase searching than others. Yahoo followed Google. MetaCrawler and Savvy Search do not or poorly support Boolean searching. Most of the Search engines had no ability to restrict search to a specific time frame. The search engines that supported gave an option to restrict time from the present going backwards to past years. The result displays were excellent for Google followed by Yahoo. Google and Yahoo are therefore recommended to give them higher priority in selecting them until they prove otherwise. However it should be noted it is not a grantee to find what you need. Therefore try as many search engines as you can to meet your information need. This recommendation is true for Agricultural subjects of which were used in the experiment.

Most of the academic or research staff and postgraduate students (92.3%) use the Internet in Literature Search. Hence Internet has become an indispensable service to the university communities, but how effectively and efficiently do users use the Internet? The question remains unanswered. Although it was revealed that most of the respondents use the Internet in Literature Search, relatively low percent (72.5%) of them had ever heard of search Engines. This is a challenge to the information professionals. The situation calls for a review of methods used to enrich the university communities with information searching skills.72.5% had implicitly no any query formulation skills. The implication is that there is a high possibility that information users spend a lot of time searching for literature with limited relevant results. The evidence is presented in table 3

where it was found that 72.5% of the respondents acknowledged that they did not know how to formulate queries, only 27.5% said to know phrase and possibly Boolean query formulation.

Yahoo was found to be the most popular as it had 67% of the respondents. Google was the second; it had 49.5% of the respondents. AltaVista ranked in the highest three with 44% percent of the respondents. As far as preference is concerned Yahoo (37.4%) was the most preferred search engine. This was the highest percent. Google was the second with 20.9%. AltaVista being the third had 8.8% of respondents. The rest of search engines had a range of 1.1% - 2.2% respondents. 16 respondents did not respond to this question.

5.2.0 Recommendations

5.2.1 More studies are recommended especially to evaluate and make more comparisons between metasearch engines and single search engines under slow connection speeds. Relatively long period studies with more search engines, more search queries from different subject areas and more connection speeds are highly recommended. This follows the fact that many books and papers have recommended metasearch engines to be a solution on limited coverage exhibited by any search engines. Inspite of this fact, there is an indication that these search engines are still subjected to timeout in slow speed connections because even when relatively long periods were specified the performance did not change. Furthermore they never improved the results in this study that is they indexed and retrieved similarly low number of documents. MetaCrawler for example was not able to index and retrieve any document at the University for Dar es Salaam Library. The recommendations to the information users and professionals in this situation are cautioned not to depend on them because they might not provide the

desired results). It is also recommended to improve the connection speed or rather the Local area network set up to allow the use of metasearch engines in our literature search.

5.2.2 Since a good number of information users do not know or have limited searching skills, the implication is that there is a high possibility that information users spend a lot of time searching for literature with limited relevant results. Information professionals have to invent or introduce basic information searching skills training to the academician to be able to match with the technological development in searching the literature. Not only to the academicians/research University staff but also University Students and if possible to think of introducing similar programmes to the secondary schools. Searching for literature on the Internet will still continue to require more than one search engine because the coverage of each one is still limited as summarised by Lawrence and Giles in 1998 and 1999.

Bibliography

1. Augustino, D. M (2000) *Examination of us of information technology application for on line searching.* A thesis submitted at the University of Dar es salaam as partial fulfilment of Masters of Arts in information studies at the University of Dar es Salaam

2. Bar-Ilan, J (2000) *Evaluating the stability of the search tools Hotpot and Snap: A case study.* Online Information Review Vol. 24 No. 6

3. Becket, D (1998) *Search Engine Corner.* http://www.ariadne.ac.uk/issue16/search

4. Chatfield, C (1983) *Statistics for technology.* A cause in applied statistics. 3rd ed. Chapman and Hall, New York.

5. Chu, H and Rosenthal, M (1996) *Search engines for the World Wide Web : a comparative study and evaluation methodology ASIS96,* http://www.asis.org/annual_96/electronic_proceedings/chu.html

6. Curtin (2001) Robot - Driven Search Engines. http://www.curtin.edu.au/curtin/1...gwpersonal/senginestudy/zseng.htm

7. De Smet, E and Kerstens, V (2000) *Supply of Academic publications (SAP): Towards a sustainable electronic document delivery system for southern libraries using the Internet.* A report for SAP project.

8. Dong, X and Su, L.T (1997) *Search engines on the World Wide Web and information retrieval from the Internet:* a review and evaluation. Online and CD-ROM review. Vol. 21 no.12 pp.67-81

9. Hanka, R and Fuka, K (200) *Information overload and "Just in time" knowledge.* Electronic Library Vol. 18 No. 4 pp 279-284

10. Google (2001) *Google.* http//www.google.com/help/

11. Introna, L and Nissenbaum, H (2000) *shaping the Web: why the politics of search engines matters.* Information Society Vol. 16 no. 3 pp. 167-185

12. Kanawha Public Library (KPL) (200) *Internet Search engines and Directories* http://kanawha.lib.wv.us

13. Katundu, D. R (1998) *The use and sustainability of Information Technology in academic and research libraries in Tanzania.* A thesis submitted as a fulfilment of PhD

14. Kibirige, H. M and De Palo, L. (2000) *The Internet as a source of Academic Research Information: Finding of two pilot studies.* Information Technology and Libraries. Vol. 19 no. 1 p. 11-16

15. Lawrence, S and Giles, C. L (1998) *Searching the World Wide Web.* Science vol. 280 pp. 98-100.

16. Lawrence, S (2001) *Online or invisible?* Nature vol.411, no. 6837, pp. 521.

17. Lebedev, A (1997) *Best Search engines for finding scientific information in the web.* http://www.chem.msu.su/eng/compason.html

18. Lesser, V. et al (2000) *BIG: an agent for resource-bounded information gathering and decision-making.* Artificial Intelligence Vol. 118 no. 1-2 pp197-244

19. Levey, L.A (2000) *Wired for Information: putting the Internet to good use in Africa.* A Project for Information Access and connectivity(PIAC). Nairobi

20. Mann, P. S. (1995) *introductory statistics.* 2nd Edition. John Wiley and sons, Inc., New York.

21. MetaCrawler (2001) *MetaCrawler fact sheet.* http://www.metacrawler.com/

22. Mnyani, Z.V.G (2000) *Factors influencing effective application and use of information and communication technology in Tanzania.* A dissertation submitted at the University of Dar es salaam as a partial fulfilment of masters of arts in information studies

23. Niewenhuysen, P. (1999) *Subjects for research or study.* Vrije universities Brussels. Brussels.

24. Northern Light Technologies Inc. (NLTI) (2001) *Northern Light Search Help.* http://www.northernlight.com/

25. Proctor, L (1997) *Academic Research on the Internet.* http://web.uvic.ca/comped/online/research/

26. Repman, J and Carlson, R. D (1999) *surviving the storm: Using Metasearch Engines Effectively.* Computers in Libraries. Vol. 19 no 5

27. Ruh, J.F (1996) *Introduction to Internet.* Eastman Kodak Company, New York

28. Salton, G (1989) *Automatic Text processing.* Addison-Wesley. Reading, M.A

29. Sasikala, C and Patnaik, K.R (1999) *A comparative study of two web search engines: Altavista and Excite.* Proceedings of the sixth national convention for automation of libraries in education and research, Nagpur. Pp 346-54

30. Scoville, R (1996) *Special reports: Find it on the net!* PC world, also available at http://www.lycos.com

31. Tanenbaum, A.S (1996) *Computer networks* 3rd Ed. Printice-Hall International, In NewJerssey

32. Tweve, J. T (2000) *An investigation of the availability and application of Information technology in the institute of higher learning.* A dissertation submitted at the University of Dar es salaam as partial fulfilment of masters of arts in information studies.

33. Vidmar, D.J. (1999) *Darivin of the web: The evolution of search tools.* Computers in Libraries, vol.19, No.5, pp 23 - 29

34. Weimer, M.L and Rusch, P.F (1996) *New searching technologies and interface online information 96.* Proceedings of the twentieth International online information meeting. pp221-4

35. Wema, E.F (2000) *The impact of introducing computers into library services.* A dissertation submitted at the University of Dar es salaam as partial fulfilment of masters of Arts in information studies.

36. Westera, G (1996) *Robotic Driven Search engines.* http://www.curtin.edu.au/curtin/l... gwpersonal/segninestudy/zseng.html

APPENDECES

Appendix A: Data Collection Sheets

Experimental Tests

This sheet was used to collect the data for the experiments

Name of the metasearch engine_____

Sheet 1: Number of Documents Retrieved

Queries	Number of documents per test											
	Test1		Test2		Test3		Test4		Test5		mean	
	T	R	T	R	T	R	T	R	T	R	T	R
Query1												
Query2												
Query3												
Query4												
Query5												

Key: T= Total number of indexed documents R= number of relevant documents

Sheet 2: Responsiveness

Name of metasearch engine_____

Queries	Search engine response time per test					
	Test1	Test2	Test3	Test4	Test5	mean
Query1						
Query2						
Query3						
Query4						
Query5						

Advanced search capabilities (features and performance):

_____

_____

_____

Capability to restrict search to a specific time frame:

_____

_____

_____

Appendix B: The Information User Questionnaire

1. Personal information

   i. Name(*optional*) _____

   ii. Occupation (*circle appropriately*) (a) University academic/research staff

   (b) Postgraduate student

   iii. Level of education_____

   iv. Sex_____

   v. Institution _____

   vi. Department _____

2. (a) Do you use the Internet in your literature search?(please circle)

   i. Yes

   ii. No

   (b) If you have answered yes in (a) above how do you gauge your frequency of using the Internet

   i. Not frequently

   ii. Frequently

   iii. Very frequently

   (c) 1) Have you heard of Internet search engines?

   i. Yes

   ii. No

   2) If yes in 1 (c) 1) above what is their role? _____

   If you know query formulation or the kind of query they accept please explain_____

   _____

   _____

   (d) Which search engines (tools) do you usually use to meet your information need? (*Please circle)*

   i. AltaVista

   ii. HotBot

iii. Snap

iv. Web Crawler

v. MetaCrawler

vi. SavvySearch

vii. Northern Light

viii. Yahoo

ix. Lycos

x. InfoSeek

xi. Google

xii. Excite

xiii. Open text

xiv. Others(specify)_____

(e) Which of the search engines do you prefer most? _____

Please give reasons: -

i) _____

ii) _____

iii) _____

iv) _____

v) _____

3. Do you usually get what you need whenever you use the Internet (using the search engine you mentioned above)? (*you can choose more than one response)*

i. Yes, I do without difficulties

ii. Yes, I do with difficulties

iii. Yes, I do, but with a lot of irrelevant materials

iv. Yes, I do with very limited search results

v. No

vi. Others(*specify*)_____

_

Appendix C: Selected and Evaluated Search Engines

Alta Vista

Alta Vista was developed at Digital Research Laboratories in Palo Alto, California and formally delivered to the Web on December 15, 1995. By January 1996 it was known to index full text of over 16.000.000 Web pages (Chu and Rosenthal 1996). In its documentation, there is an indication that AltaVista can fetch2.5 million pages a day following the robots Exclusion standard, and index IGB of text per hour.

Its search tips for receiving more precise results includes the following recommendations: use an exact phrase, specification of the language, use lowercase text in searches because the search service finds both uppercase and lowercase. Using uppercase text it only finds uppercase results. Other tips are for include or exclude words. Include words by placing a plus sign (+) before the keyword. Exclude by placing a minus sign (-) immediately before the keyword. Use wildcards by typing and asterisk (*) at the end of a keyword to search multiple forms of words e.g. big*, to find big, bigger, biggest and bigwig.

AltaVista has both simple and Advanced search services, others are images, MP3/Audio, video chips, Web Directory. In the advanced search you can specify time frame, language, URL and number of results per page. Also there is an option of text search which can find results from UK, URLs pr World wide similarly the text search contains simple and advanced search

Excite (http://www.excite.com)

Excite claims to cover all the World Wide Web pages, news, maps, yellow pages, stock quotes, TV listing, weather, E-mail addresses air line flights (Sasikala and Patnaik, 1999). The searching mechanisms include simple and advances search options. The

general search Tips include: use more than word and be specific, choosing a specific index, use quotation for searching an exact phrase and uses plus (+) sign to obligatory include the word and minus (-) sign to exclude the unwanted word in the search results.

The advanced search option includes use of Boolean operators - allows concept based search mechanism to turn off, allowing the user to search for documents that contain exactly the words being searched. Boolean operators include AND, AND NOT, OR and parentheses. They must appear in ALL CAPS with space on each side.

Search results consist of "web results" show Titles and view by URL. Excite lists ten search results at a time. It claims to list the most relevant document first. Show titles option displays titles and URLs of the results. View by URL displays only the names of the URLs and the relevant links with them. This search engine was developed by Architext software. Scoville (1996) reported that excite claims 1.5 million fully indexed Web pages.

Google (http://www.google.com)

Google claims to index all of the World Wide Web (Google, 2001). It is a search engine found to have less information in the published research literature implying that possibly few researchers have researched on it. Its documentation reveals that a query should have few descriptive words. Google uses sophisticated text matching techniques to find pages that are both important and relevant to the information search. It assigns higher relevance to pages in which query terms appear near each other. Google has automatic "AND" operator. This means it does not require the user to add this operator. It returns pages that include all of the search terms. The "OR" operator is supported to retrieve pages that contain either word A or word B. Common words and characters (known stop words) are ignored, because they slow down searches without improving the quality of results. Examples of stop words are "what", "how", "a", and "the".

A plus sign (+) can be used if the word must be included in search results. Words can be excluded by typing minus sign (-) before the word. Search terms in bold can be viewed in the results to enable the user see at glance. Google does not use "stemming" or support "wildcard" searches. It searches for exactly the words entered in the search box. It is further not case sensitive. All letters no matter how they are typed are understood as lower case. Phrase searching is supported. For exact phrase double quotes are used hyphens, slashes, periods, equal signs and apostrophes are recognised as phrase connectors and these work like the quotation marks. With Google search can be useful to restrict to a web site. Advanced Search option is available. By simply adding more words to a broad query often narrows it leading to get what is wanted. Advances operators include "link:" "OR", "+ ", "-".

Hotbot (www.hotbot.com)

Hotbot is a successor of the inktomi search engine. Paul Gauther and Eric Brewer developed it at the University of California, Barkeley (Westera 1996). It was released in 1996. Hotbot has both simple and advanced searching mechanism, use of double quotes support for an exact phrase to be included in the search results. Plus or minus signs (+/-) may be used by placing immediately in front of search terms implying that such particular keyword will definitely appear in the search results. Simple Boolean operators "AND", "OR", and "NOT" are supported. Proximity searching is not supported (Westera 1996). It returns a list of hyperlinks with a short description of each found document (Lebedev, 1997).

Infoseek (www.infoseek.com)

Infoseek Corporation developed this search engine. It became available as a beta version in August 1996 (Westera 1996). Infoseek claims to search for a word, group of words or a phrase. During its search takes account of different word forms such as singular and plural forms (Lebedev, 1997). Time frame specification is also not present.

It supports plus (+) and minus (-) that must be placed immediately in front of the word sign to indicate that the word must be included. A quotation is used to delimit a phrase. Results are displayed such that there is a list of hyperlinks with a short abstract for each document.

MetaCrawler (www.metacrawler.com)

MetaCrawler, part of InfoSpace Inc., was developed in 1994 by then University of Washington (UW) graduate student. This is one of the first metasearch engines. MetaCrawler is said to be one of an effective and easy to use metasearch tool (Repman and Carlson 2000). It queries many of the web's top search engines simultaneously. It draws upon the databases of large number of the webs best search engines. MetaCrawler simultaneously sends queries to AltaVista, Excite, Google, GoTo, Looksmart, Lycos, MetaCatalog and WebCrawler, among others.

In the literature it is said that many industry publications and analysts have recognised MetaCrawler as the Internet's finest search service. This includes "Best Search Engine" twice by PC Magazine, determined to search more than twice as much of the Web compared to its nearest competitor by a 1999 NEC Research Institute study (MetaCrawlwer, 2001). Most recently, Yahoo! Internet Life (July 2000) named MetaCrawler one of the 10 "Internet essentials".

Northernlight (www.northernlight.com)

This search engine was developed and released in 1997 (NLTI, 2001). Northernlight demands to index the whole World Wide Web. Its searching mechanism consist of simple search and power search implied Boolean operators are supported such as plus (+) and minus (-) sign which are supposed to be placed in front of a search term leaving

a space between the sign and the term that should be included or excluded respectively. Simple Boolean operators "AND", "OR", and "NOT" are also supported. Proximity searching is not generally supported. Power search is documented as a complete toolkit for advanced searching by subject, source, document type, and date. These means limits may be made to subjects document types and date of publication. It further claims to support natural language searching. The more words used the more on-target the results are returned. For example ergonomic workstation mouse keyboard instead of ergonomics.

Two kinds of truncation symbols (wildcards) in queries are supported. The * (asterisk) to replace multiple characters and the % (percent) symbol to replace only one character. Northernlight automatically stems most common plural and singular forms of words. A search on "cat" will also return results containing the word cats and a search on cats will return results containing the word cat. Displayed search results always may have search terms not in bold.

SavvySearch (www.savvysearch.com)

SavvySearch is one of the metasearch engines. It claims to serve as an access point for over 200 search engines (Repman & Carlson 1999). It has default Boolean AND search and selecting a phrase search is possible. Also it uses OR and other operators such as plus (+) sign and minus (-) sign. The user doesn't have control of the number of links obtained or the wait time. Results are displayed that comprise a summary with attribution. According to its documentation searching once you may get results from over 1,000 search engines, web directories, auction, storefronts, news sources, discussion groups, reference sites, and more. The search sends query to several search engines at one time and integrates the results into one list. It may supports advances search options:

- double-quoted phrases:"" (e.g., "john lennon" not john lennon)

- enforced term operators: +/- (e.g., music + "john lennon" -beatles)
- Boolean language: and , or, not (e.g. music and " john lennon and not beatles)

WebCrawler (http://www.webcrawler.com)

Brian Pinkerton developed WebCrawler at the University of Washington in 1994. American Online took control of it mid 1995. WebCrawler claims to index all the World Wide Web. Through its partnership with LookSmart. WebCrawler claims to offer over 1.5 million websites and 100,000 organised categories of sites directory. Its database is said to be small (Curtin 2001)

Its documentation indicates that its search technology analyses the documents in the WebCrawler search index to select the most relevant and popular web pages to match any query. Its interface offers option to search for web sites, News, photos and all. The search tips include using more than one word, choosing a specific index i.e. web or all or News, or photos, use of quotations, use of plus sign minus and Boolean operators (AND, AND NOT, OR and parentheses).

Yahoo (http://www.yahoo.com)

Yahoo documentation (2001) reveals that, two men: David Filo, and Jerry Yang while attending Stanford University developed yahoo. Yahoo is an abbreviation for "Yet Another Hierarchical Officious Oracle". In the documentation it claims that the acronym represents the fact that Yahoo seeks to be a directory or hierarchy that serves as an oracle or knowledge giver to the modern day office dweller who is officious.

Yahoo searching mechanisms include simple and advanced search. The later supports phrase and Boolean searching. The searching tips include use of double quotes between

phrases, use of plus sign (+) and minus sign (-) if the key word must or must not appear respectively. The advanced search uses in-built Boolean operators specified by using radio buttons. The option include intelligent default, exact phrase match all words (AND) and match on any word (OR). Time frame restriction is possible to refine literature or information added in the past four (4) years. Google powers yahoo. These results don't come only from Yahoo! Index but also from the index of search partner, Google. The results are displayed in bolds the search terms entered the search box.