OPTIMISING INFORMATION RETRIEVAL FROM THE WEB IN LOW-BANDWIDTH ENVIRONMENTS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE,
FACULTY OF SCIENCE
AT THE UNIVERSITY OF CAPE TOWN
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTERS IN PHILOSOPHY (IN INFORMATION TECHNOLOGY)

BY
ASHWINKOOMARSING BALLUCK

FEBRUARY 2007

Supervised by Hussein Suleman

ABSTRACT

The Internet has potential to deliver information to Web users that have no other way of getting to those resources. However, information on the Web is scattered without any proper semantics for classifying them and thus this makes information discovery difficult. Thus, to ease the querying of this huge bin of information, developers have built tools amongst which are the search engines and Web directories. However, for these tools to give optimal results, two factors need to be given due importance: the users' ability to use these tools and the bandwidth that is present in these environments.

Unfortunately, after an initial study, none of these two factors were present in Mauritius where low bandwidth prevails. Hence, this study helps us get a better idea of how users use the search tools. To achieve this, we designed a survey where Web users were asked about their skills in using search tools. Then, a jump page using the search boxes of different search engines was developed to provide directed guidance for effective searching in low bandwidth environments. We then conducted a further evaluation, using a sample of users to see if there were any changes in the way users access the search tools.

The results from this study were then examined. We noticed that the users were initially unaware about the specificities of the different search tools thus preventing efficient use. However, during the survey, they were educated on how to use those tools and this was fruitful when a further evaluation was performed. Hence the efficient use of the search tools helped in reducing the traffic flow in low bandwidth environments.

CONTENTS

CHAPTER 1

1.0 Introduction ……..1

CHAPTER 2 – Background

2.0 Information Retrieval and its needs……4

2.1 The World Wide Web (WWW)……5

2.2 Information retrieval & the World Wide Web……6

2.3 Information Discovery……...8

2.4 Search engines……....8

2.5 Meta search engines ……11

2.6 Web directories……...12

2.7 Newsgroups and Mailing Lists……13

2.8 Problem arising when using information Retrieval Tools and its optimisation……..15

2.9 Mauritian Infrastructure……16

CHAPTER 3 – Determining User Habits

3.0 Pilot Study……20

3.1 Description of Survey……...21

3.2 Sample Population ……22

3.3 Findings of Survey

3.3.1 Demographics……..23

3.3.2 Discussion of Responses……24

3.4 Summary of Findings……...43

CHAPTER 4 – Bandwidth Optimisation Tool

4.1 Overview……44

4.2 Further Evaluation……..46

CHAPTER 5

5.1 Conclusion……...50

5.2 Future Work……51

BIBLIOGRAPHY……52

APPENDIX

CHAPTER 1

1.0 Introduction

The enormous amount of digital information accessible on the Web today poses a great challenge to users who need to look for specific information. Thus, to overcome such problems, many tools such as search engines and Web directories have been created to ease the retrieval of information from this large bin of unclassified information. However, these tools as developed do not adequately solve the problems faced by Web surfers. The main reason for such difficulty is that the users are not well informed and they do not possess the required skills to make maximum use of these tools.

The aim of this study is to investigate the habits of users in terms of the use of such tools with the help of a survey and subsequently evaluate the improvement in users' ability to locate information using tools that highlight advanced features. The tools include some popular search engines, Web directories and mailing lists. Furthermore, a jump page which includes the search boxes of the different search tools was developed and the users were further evaluated using this tool.

This study will be useful since in countries like Mauritius where a low bandwidth environment prevails, it becomes important to optimise bandwidth use. Hence, by educating users on the proper use of information retrieval tools, unnecessary traffic flow that normally occurs when users retrieve useless information during typical search sessions could be reduced.

As earlier indicated, the tools will include search engines, Web directories and mailing lists. Moreover, users can retrieve information from the Web through two methods and these are known as the full text search and the hypertext system. The full text search is performed through the use of search engines and the hypertext system is analogous to browsing the Web. These two methods are known as the pull systems where the users' interaction needs to be good to get fruitful information. We also have the push system structure

where the Web pushes information to the users. Such type of information gathering can be achieved by the use of mailing lists.

Both architectures will be covered during this study. The pull system will be covered through the study of the behaviour of users with search engines and Web directories and the push system will also be taken into consideration even though user interaction is not needed. The reason is that with the push system, broad topics of information are sent when the user sends a request and thus, as the topic needs to be modified, new requests need to be sent and thus this increases the bandwidth usage. However, for the mailing lists, even though user interaction is not needed, it will be covered in this study since unnecessary information might be sent to users who have once subscribed to those mailing lists. Thus, in summary, the pull system allows the user to change the search topics according to the requirements each and every time a particular information is needed whereas for the push system, information is constantly being sent although it might not be wanted.

The main question that is asked while trying to investigate how useful a search engine is to users is: 'how is a person using a system?' This question leads us to the point that designers of information retrieval tools do not adequately plan the Graphical User Interface of search engines. Instead they rather try to optimise the back end algorithms of the search tools. However, to optimise the use of search tools, the designers as well as the researchers need to figure out what the needs of the individual searcher are or how behaviour differs for different tools and develop search tools according to their interactions.

Until now, most of the research done to evaluate users' information retrieval skills have been through "large sets of user-system transactional data taken from real online searches on the Web" (Martzoukou, K. 2005) and this type of evaluation is considered to be the most reasonable and non-intrusive means of collecting user searching information from a large number of users (Jansen, B.J. & Pooch, U. 2000). However, there is another well-known type of evaluation that may provide fruitful information and this is through surveys and

questionnaires and these are based more on hypothetical searching rather than empirical situations.

Research on the Web gives information that is “useful for examining behaviours and actions but is not adequate for explaining the factors and processes that have led to that behaviour” (Martzoukou, K. 2005). We may thus deduce that user knowledge is directly proportional to the web performance of information retrieval tools. Thus, in this research, it is attempted to show that poor user interaction with information retrieval tools occurs due to lack of knowledge. Hence, the results of this study may apply to the above statement of Martzoukou.

CHAPTER 2

Background

2.0 Information retrieval and its needs

The term ‘information’ has a high value in our everyday lives and thus they need to be well kept for future reference. However, retrieving relevant information from any database remains a challenge. Thus, due to this most difficult task, many researchers have tried to develop general and efficient strategies for information retrieval. As defined by Baeza-Yates and Ribeiro-Neto (1999), “information retrieval deals with the representation, storage, organisation and access to information items”. The representation and organisation of information items should provide the user with easy access to the information in which he is interested. Moreover, Zhang et al. (2005) stated that due to “the enormous amount of information accessible today, it poses a great challenge to information retrieval systems to effectively retrieve information to satisfy the users’ needs”.

In theory, information storage and retrieval is simple. Considering an example where documents are kept in a store e.g., the Web, when information needs to be retrieved, a person needs to formulate some queries in order to get the required information by discarding useless items. This is known as a ‘perfect retrieval’.

In practice, it becomes really difficult to retrieve required information from a huge repository of information. In natural language processing, it is impossible to characterize free text completely due to the difficulty machines have in understanding human languages. Thus, in order to overcome this, information retrieval tools need to be built using statistical information as well.

For a better understanding of the characterization of documents, the concept of information retrieval vs. data retrieval needs to be explained. While speaking of information retrieval, most people tend to mix the concepts of information retrieval with that of data retrieval. Data retrieval aims at retrieving

all objects which exactly satisfy well defined semantics and conditions e.g., the list of all oceans. On the other hand, information retrieval deals with imprecise retrieval of objects which might be useful to a particular user.

Van Rijsbergen[1] in his book differentiated information retrieval from data and he tabulated its properties as follows:

	Data Retrieval	**Information retrieval**
Matching	Exact match	Partial match, best match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Classification	Monothetic	Polythetic
Query language	Artificial	Natural
Query Specification	Complete	Incomplete
Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive

Table 1. Data Retrieval v/s Information Retrieval

Thus, as we can deduce from the properties in Table 1, the relevance of the document is high in a data retrieval system. But, in the case of information retrieval, the relevance might vary. Thus, this notion of relevance is at the centre of information retrieval. In summary, as stated by Baeza-Yates and Ribeiro-Neto (1999), “the primary goal of an information retrieval system is to retrieve all documents which are relevant to a user query while retrieving as few non relevant documents as possible”.

2.1 The World Wide Web (WWW)

The World Wide Web (WWW) is a very large distributed information space. Its origin was in CERN in 1991 where research documents in nuclear physics were shared in an organization-wide collaborative environment. Nowadays, the WWW has grown to provide lots of information such as homepages,

[1] Information Retrieval www.dcs.gla.ac.uk/keith/Preface.html

publications, libraries and general information. The WWW can be defined to be a system of Internet servers that support specially formatted documents. The documents are formatted in a markup language called HTML (HyperText Markup Language) that supports links to other document, as well as graphics, audio and video files[2]**.** However, the classification of document is really difficult on the web and as defined by Shen et al., "Web-page classification is much more difficult than pure-text classification due to a large variety of noisy information embedded in Web pages" (Shen, D. et al. 2004).

A study by Inktomi and NEC Research has shown that there exists around 2 billion websites and this number is to increase (Campbell, K. 2000). It has always been impossible to estimate the number of websites since, information is being added to the WWW constantly. Thus, to retrieve information from that large repository, we need to use information retrieval tools. Moreover, there exist a lot of strategies and technologies to enable the retrieval of information and this is a part of this study.

2.2 Information retrieval and World Wide Web

As elaborated in the section above, the Web is a large bin of information. Thus, the ability to search from the Web efficiently and effectively is a technology by itself. As noted by Martzoukou, "the Web has grown into a vital channel of communication and an important vehicle for information dissemination and retrieval" (Martzoukou, K. 2005). Many users think that querying information from the Web can be done at only a mouse click, which seems reasonable, but in fact the situation is more complex than that. As described by Gordon & Pathak (1999) "the primary use of the Internet is for information retrieval; but due to its complexity, information retrieval on the Web is really difficult and cannot be compared to other forms of information retrieval". The reason for such problems is that the WWW is a huge library of information that is heterogeneous and dynamic and contains multimedia

[2] www.webopedia.com

elements. Also, there exist a lot of hyperlinks from Web pages that lead to useless information or spam.

Moreover users, especially novice and irregular users, find it difficult to phrase their information needs due to the lack of knowledge and literacy in search domains. (Bates, M.J. 1998). Information overload on the Web also contributes a lot to preventing Web surfers from getting access to useful information. As defined by Nelson, “information overload is the inability to extract needed knowledge from an immense quantity of information for one of many reasons” (Nelson, M. R 1997). Wurman, R.S. (1989) on his part gave a detailed explanation of how information overload occurs. He states that information overload may occur when a user:

- does not understand available information,
- feels overwhelmed by the amount of information to be understood,
- does not know if certain information exists,
- does not know where to find information, or
- know where to find information, but does not have the key to access it.

From all the above points, we can see that a user experiences many limitations while trying to retrieve information from the Web. Thus, this increases the anxiety of the user and who just abandons the process or takes references to some useless information, assuming that they may be the right ones. Information overload might be a good reason for information retrieval tools to be integrated with the Web.

We know that large volumes of uncontrolled information, such as the Web, are potentially full of errors, inconsistencies and useless data. Thus, any retrieval of such information might not yield anything useful. From Nelson’s point of view, “when we try to retrieve or search for information, we often get conflicting information which we do not want” (Nelson, M. R. 1997), thus making information retrieval a challenging problem. Information overload on the Web makes information retrieval a laborious task for any user. They will have to do several searches or refine searches before coming to a relevant and accurate document, bearing in mind that only the user who is using an

information retrieval tool can judge the relevance of the document viewed. Hence, once more, to allow easier and simpler access to millions of resources that are available but hard to find, information retrieval tools need to be integrated with the Web.

2.3 Information Discovery

In this study, we have been speaking about information retrieval tools that help users to retrieve useful information from the Web. These tools are numerous and manipulations of each of each such tool differ from others. As described by Gray (2004), “using the various search tools on the Web is enhanced by knowing how they were actually designed, and especially by knowing the specific rules, all too often different for each tool”.

Accessing information on the Web can be achieved through different ways and these are listed below.

- Going directly to a site if you have the address.
- Browsing.
- Exploring a subject directory.
- Conducting a search using a Web search engine.
- Exploring information stored in online databases on the Web, also known as the ‘Deep Web’.
- Joining an email discussion group or Usenet group.

However, as elaborated before, this study will cover mostly search engines, Web directories and mailing lists.

2.4 Search engines

“Search engines are considered to be the most important tool for retrieving information on the Web and, consequently, form a critical area of research“(Gaines, B.R. et al., 1997). Usually search engines are useful in querying information from the large Web space. These databases may

contain any type of file including text, sound, graphics or even video. This tool works like a database retrieval system that searches gigabytes of indexed databases in seconds. Thus, this helps users to find specific information that is needed without much processing time.

We can categorize the Web search engines into three components: the spider, index and query mechanism components. The spider, which might be referred to as a robot, walker or crawler is a program that traverses the Web from link to link, identifies and reads the pages. The indexer then figures out which words to exclude from an index; searches through a list and finally builds an index that contains relevant associations between remaining words and documents. Since most indexers build a full text index based on the documents, the index is often larger than the original files.

“Web Servers and clients use the client-server paradigm to communicate” (Gudivada et al. 1997). Using this architecture, robots find it easy to traverse the Web efficiently. Generally, there exists three traversal methods and these are:

(i) The robot is provided an initial URL. This URL is indexed by the robot which later extracts all URLs pointing to other documents and then examines each in a breadth-first or depth-first way.

(ii) The Web is partitioned by Internet names or country codes and robots are assigned to explore the space. This is normally the most common method.

(iii) Some URLs with high popularity are searched most often. Thus, there is the expectation that such types of homepage contain the most frequently sought information.

Moreover, we might see other types of indexers that use a synonym list. This helps users to find what they need on the Web without knowing the exact words. They just have to input any word and its synonyms are used interchangeably.

The search engines query mechanism is the software that enables any particular user to query the index and return the results in terms of relevancy. As described, the above three components form the general way that a user might think of the processing of a search engine. The input of the query term initiates the search process. Thus, the engine searches its index and generates a page which links to those resources containing some or all the terms and these are usually presented in a ranked order. This type of ranking could favour documents where the search terms appeared many times near the beginning of the document, close to the title or in the title of the document. The above processing described may be thought of as the first generation of search engines.

However, development in search engine technology and criteria for ordering of search results have changed. They now take into account several aspects such as keywords, sites, links and popularity. To be more explicit, the ranking of search engines in the back end depends on three factors i.e., its relevance to the words and concepts in the query, its overall popularity and whether search engine optimisation (SEO) has been performed for the websites being displayed.

The indexing process of search engines also helps in providing information efficiently. “The process begins by removing all the very frequent and non-significant words (e.g., the, are, or, of)”. (Savoy, J. 2000). Then, to assist in a more successful search, a stemmer and thesaurus is also added to the search engines where the stemmer has the ability to search on a word root that can have multiple endings. Thus, in the stemming procedure, inflectional and derivational suffixes are removed and, finally, the keywords are weighted. We might consider the TREC conferences to be one among the best weighting procedures. (Savoy, J. & Picard, J. 2001). The details of the findings from the TREC conferences are as below.

(i) If a word appears more frequently, its weight is increased.
(ii) If a word appears within many pages, its weight is decreased.
(iii) More weight is assigned to short pages than longer ones.

(iv) An inverted file is updated such that the system is able to obtain a list of all Web pages for each keyword.

We might also say that search engines perform a filtering task to get rid of duplicate sites with no importance. As described by John M Lervik, owner of FAST SE, performing a 3rd generation search "means getting close to the user's real intention, applying rules of grammar, syntax and semantics to computer linguistics." (Wylic, I. 2002).

2.5 Meta search engines

As we have described above, users tend to search the Web using information retrieval tools, among which is the search engine. However, due to the numerous search engines and their different types of processing whilst retrieving information from the Web, each of the search engines displays different results for the same keywords input for search.

Thus, in order to overcome such problems, developers have developed another tool known as the meta search engine. Meta search engines are search engines that display results from multiple search engines. "They actually do not have their own databases, do not collect Web pages, do not accept URLs additions and do not classify or review websites unlike search engines" (Liu, J. 1998). When queries are entered in the search box, the meta search engine transmits the search simultaneously to individual search engines and their databases of Web pages. Duplicate findings are however merged into one entry and some results are ranked according to some criteria in those meta search engines. There exist some systems which also allow selection of search engines to be searched.

In meta search engines, the query submitted might not be as detailed as that of search engines. Boolean searches might produce varied results; phrases and query refinement might not be supported. Moreover, searches by meta search engines are done only on part of the databases of search engines. As defined by Liu, J. (1998), "meta search engines generally do not conduct

exhaustive searches; they do not bring back all the pages from each of the individual search engines". They only make use of the top 10 to 100 hits from each of them.

2.6 Web directories

We have been discussing the different methods of retrieving information from the Web, of which one more item is the Web directory. Web directories can be defined to be a combination of searching with browsing. Looking for information in a Web directory can be achieved in two ways: browsing the directory or searching its contents. As described by Baeza-Yates, "Web directories are hierarchical taxonomies that clarify human knowledge" (Baeza-Yates R. and Ribeiro-Neto, B. 1999). We could also say that the directory is something akin to a huge reference library and people sometimes describe them as catalogs, yellow pages or subject directories.

Web directories are considered to be a better retrieval tool than search engines, although the Web coverage of all the directories is less than 1 % of all Web pages (Baeza-Yates, R. and Ribeiro-Neto, B. 1999). The reason is that the results returned from directories are far more relevant than other tools. Also, following observations and interviews on using information retrieval tools carried out by Fidel et al. (1999), students preferred searching using Web directories, relied on past successful search experience, used landmarks, performed swift and flexible searching and were generally satisfied with the results, but were impatient with slow system responses. However, this study might not hold true nowadays since Google has developed better techniques to include in their search engines and this encourages Web searchers to use Google.

Since Web directories are considered to be user friendly with flexible Graphical User Interfaces, a lot of universities, libraries, companies, organisations and individuals are creating subject directories to catalog portions of the Internet. These are organised by subject and consist of links to

the Internet resources relating to these subjects. Most of them provide a search capability that allows you to query the database as well.

When search engines and Web directories were developed, there was a lot of discussion on when to use which tools, and thus a well-defined statement was derived. If someone is interested in broad general information, the first place to go is a Web directory but, if the topic is narrow with specific information, the search engine is better. But, nowadays, there exists less difference between search engines and Web directories. An example is Yahoo, which is the best and oldest example of a Web directory (Baeza-Yates, R. and Ribeiro-Neto, B. 1999) that has been de-emphasizing that aspect towards the search engine option. Moreover, considering the case of Google search engine, a user need not choose between the Google Web search and the Google directory and the reason for this is that the Web search already indexes the Google directory.

However, one good reason for users using Web directories is that they may need to see sites that have been evaluated by an editor. Thus, this helps them to get precise information since human beings do the reviewing and classification. An example of such sites is http://dmoz.org/, (The Open Directory Project) where around 2000 volunteers do the classification.

Within the classification of the Web directory, there exists also a concept known as the 'deep' or 'invisible' Web. In such cases, information is stored in databases that are invisible to the search engines. Thus, crawlers from search engines are unable to find such information while indexing. In these circumstances, the best way to access this information is to search through these databases directly and these are published through different sites on the Web. http://invisible-Web.com/ is one such site.

2.7 Newsgroups and Mailing lists

Another feasible way of looking for information is through the use of newsgroups and mailing lists. Since we are trying to optimise the information

retrieval from the Web, the use of such tools might be important since they are considered to be of low bandwidth consumption.

Newsgroups, better known as Usenet newsgroups, are a distributed system of messages, like a worldwide chat system. Since many messages are sent everyday, they are divided into “newsgroups” with each newsgroup concentrating on one topic. A user can read the newsgroup’s articles, post replies or post new articles.

Like the email functionality, articles can be read according to the user’s schedule rather than in real-time and that particular article can also be read by any user. Newsgroups are named by using a hierarchical system of words, separated by dots. The first word of a newsgroup name indicates the top level category of the newsgroup. The following are some of the top levels in the newsgroup hierarchies:

Hierarchies	**Topic**
comp	Computer Hardware, Software, networking or any other computer related topics
misc	Miscellaneous topics
news	Usenet News
sci	Scientific Topics
soc	Social Topics
talk	General Discussion

Table 2. Top Levels in Newsgroup Hierarchies

Mailing lists generally work in the same way as newsgroups. However, mailing lists are built by using email addresses which are compiled and assigned to a list. When a message is sent to the mailing list name, it is automatically forwarded to all addresses in the list. Unlike newsgroups where anyone can read articles, mailing list users need to be subscribed to the list.

As described before, unlike the search engines and web directories which use the pull system structure, the above applications use the push system structure to provide information to users. In the pull system, constant user interaction is needed, since the users will need to key in their search topics to query the Web. However, for the push system, once the user's interest topic is known to any web application such as mailing lists, constant information about that topic is sent to the user.

2.8 Problems arising when using information retrieval tools and their optimization

We have been speaking of different tools for information retrieval. Now we consider the problems which arise when using them. Nowadays, most of the information retrieval tools have been designed for countries with high bandwidth infrastructure. However, these tools cannot be used efficiently in developing countries like Mauritius due to several reasons, amongst which the main one is the restricted bandwidth in the region.

However, there is also a lack of bandwidth-sensitive optimisations in the tools and this need to be addressed for smooth and fast processing. Any user of the search tools in low bandwidth countries will always find that the time taken for getting information is too long. Moreover, we might also find that search engine provide information that is not relevant to users' search or the link provided may no longer exist.

Retrieving information from the Web is complicated because all the entities of a search process i.e., users, bandwidth and search tools are not well optimised for certain regions of the world. Optimisation needs to be done in order to allow "users less frustration because they can understand the current system and predict system response". (Schneiderman, B. 1998)

For information retrieval, users use search tools to input keywords which in turn use bandwidth to access search engines for information retrieval. Hence, bearing in mind the way information is retrieved, all the entities at stake in this

processing need to be optimised. The users who provide queries need to be well trained since by feeding bogus or sub optimal information to the search tools, bandwidth may be wasted. Moreover, the search tools and the algorithms need also to be appropriately developed since when they interact with the Web, useful information needs to be fetched and delivered to the user as wrong information will also waste the precious bandwidth.

Hence, for better use of the bandwidth, both the front end (user interaction) and back end (search tools and its algorithms) need to be optimised. However, in this study we are considering only optimisations for the user front-end.

2.9 Mauritian Infrastructure

The Republic of Mauritius is located in the Indian Ocean about 800km east of Madagascar. It comprises of 4 islands Mauritius, Rodrigues Is., Agalega Is. and Cargados & Carajos Shoals of which Mauritius is the largest and is being considered in this study. It has a total landmass of 2040 sq. km and an estimated population in 2006 of 1,240,827 (CIA, 2007). The population of Mauritius is evenly distributed all over the island except for the Western and South Western area where the population settlements are low.

The Internet connection was introduced in Mauritius in January 1996 with a 128Kbps connection and in February of the same year, Telecom Plus, a joint venture between Mauritius Telecom and France Telecom (ITU, 2004), started commercialising the Internet. The introduction of Internet has provided unlimited resources to the world but in Mauritius the change was not significant since the exposure to the web was low. The individual household has been using dial up 56Kbps connections for accessing the Web since the beginning. However, in 2001, the monopoly of Telecom plus as an ISP was removed and other key players came in the market as ISPs. This helped in improving the network connections on the market and network connections with speed of 512Kbps and 1Mbps were introduced.

Although these bandwidths were introduced in Mauritius, the costs of enjoying these services are high and not affordable by individual households. The reason for such high prices is that the monopoly of the SAFE cable is still being held by Mauritius Telecom which charges a high rate to ISPs thus leading to a direct impact on the customers' pricing. For a comparison, Internet access cost in Mauritius is nearly 3 times higher than its neighbouring island, Reunion Is. (AfrISPA, 2005).

The following companies hold a licence as Internet Service providers in Mauritius (AfrISPA, 2005) and all of them need to connect through Mauritius Telecom Ltd, the holding company of the ISP, Telecom plus Ltd.

- Africa Bridges Networks Ltd
- City Call Ltd
- Clusterway Ltd
- Data Communications Ltd
- Emtel Ltd
- Harel Mallac & Co. Ltd
- I-Telecom Ltd
- Mauritius Post Net Ltd
- MFDC Ltd
- Network Plus Limited
- Paging Services Ltd
- Rogers Telecom Ltd
- SITA
- Telecom Plus Ltd

Although the list of ISPs in Mauritius is long, it does not help in improving the Internet connection which is provided to the general public. The reason is that only 4 companies are currently providing constant and reliable connections to the general public; Data Communications Ltd, Mauritius Post Net Ltd, Network Plus Ltd and Telecom Plus Ltd. From the remaining companies, MFDC Ltd, Rogers Telecom Ltd, Harel Mallac & Co. Ltd and SITA are classified as private ISPs and there services are enjoyed only by internal clients (AfrsISPA,

2005), Emtel Ltd provide Internet connection on his mobile network and the remaining on the list are somewhat inactive in the public market.

The latest evolution in Internet connection is the introduction of wireless connections of 128Kbps and 512 Kbps. The wireless infrastructure is explained in the two figures[3] below. The first figure shows the main server which distributes wireless connections to Web users in Mauritius.

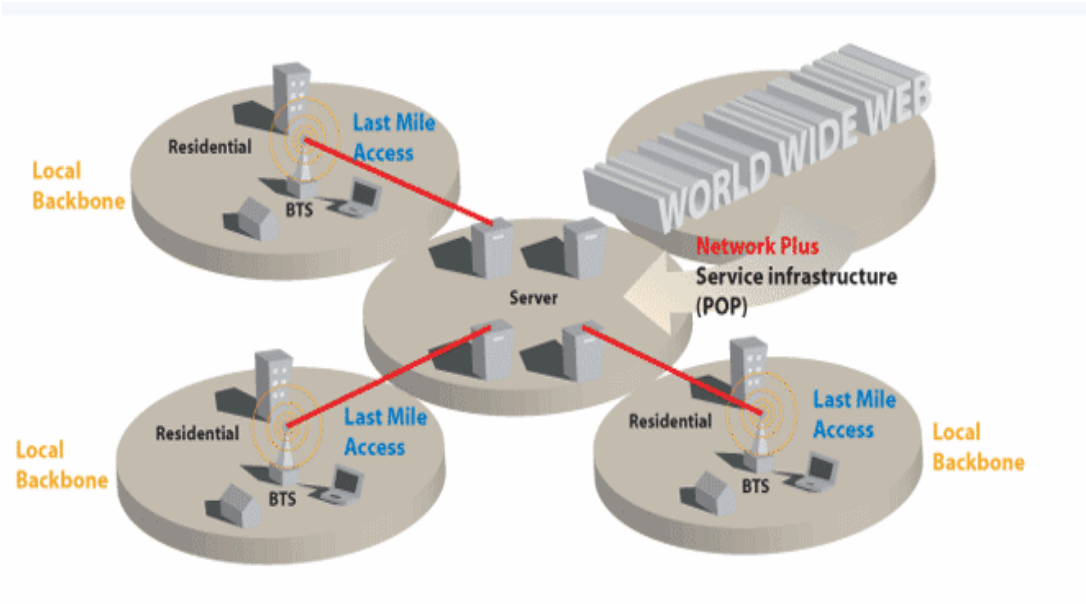


Figure 1. Wireless infrastructure of Mauritius

This service was introduced in Mauritius in 2005 by Network Plus and has been named Nomad. However, this connection has some constraints since only one third of the island is covered. This is shown in the figure 2 below.

[3] www.networkplus.mu

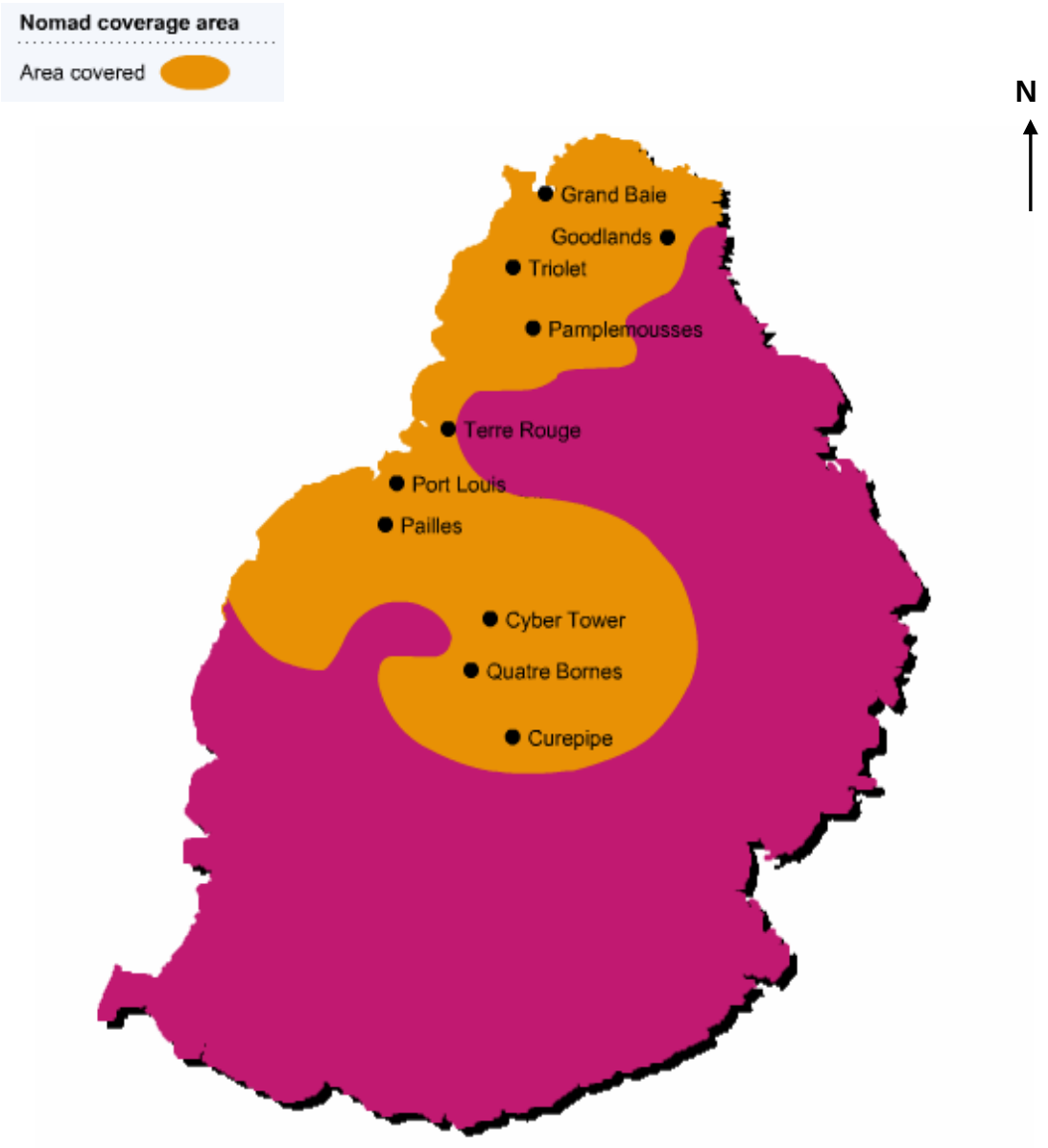


Figure 2. Wireless coverage of Mauritius

Hence, we can see that under these conditions, almost two third of the Internet users will not enjoy high bandwidth since the population is evenly spread over the island. Thus, congestion will still exist when retrieving documents or information from the Web.

CHAPTER 3

Determining User Habits

3.0 Pilot Study

As already discussed in the background section, this study began because the retrieval of information from the Web in African countries where low bandwidth prevails is a big problem. In view of solving this, we have been trying to optimise the use of information retrieval tools from the user-end i.e., front-end. However, due to lack of information from sources available, it has been difficult to determine users' behaviour with those information retrieval tools. Thus, usage patterns of information retrieval tools needed to be determined.

The first step for such investigation was to get the users' knowledge. A quick interview was carried out among 5 users. They were asked to search for document containing the "Mauritian National Anthem" and then the following questions were asked:

1. If ever you look for information on the Web, how do you proceed?

2. What are the different search engines you use? Describe why.

3. Do the search engines meet your requirements? Why?

4. Is your processing of information done in a reasonable time?

5. Do you think that there exists ways to improve the speed of information retrieval? How?

6. Why do you think a toolbar is provided with most search engines? Do you use them?

7. Do you use the 'find similar pages' in the results display?

8. Is there any other way you think you may optimize the search engine?

The users seemed a bit lost and were unable to respond to most questions. Moreover, after the compilation of results from this small demonstration and questions, it was hypothesized that all the users had a single particular way of searching the Web. Their first operation was to open the search engine homepage and input keywords in the textbox provided. If ever the first page did not retrieve any useful information, they just clicked on the back button and then re-inputted other keywords. This was being done until they found a link on the first page which might attract their attention. Moreover, the users were unaware that while performing these actions, they were wasting precious bandwidth since they were querying the international Web server several times.

3.1 Description of Survey

To better understand the habits of users with information retrieval tools, I then used written user surveys which are considered to be a familiar, inexpensive and generally acceptable companion for usability testing and expert reviews. (Schneiderman, B. 1998). The use of such surveys is practical since it allows any surveyor to cover a wide range of users. The users can vary between novice users and expert reviewers.

Moreover, as described by Schneiderman (1998), “a survey should be well prepared, reviewed among colleagues and tested with a small sample of users before a large-scale survey is conducted”. All the above conditions were given due importance before the preparation of this survey.

Any survey that is carried out for information retrieval tools might be classified into three levels: Micro level, Middle level or Macro level. Experience in using information retrieval tools should be given the highest level of importance in such a survey. However, the level of experience in interacting with the World Wide Web (WWW) and also the familiarity with computers - the degree to which a user understands a computer and its limitations and advantages - influence the searching strategies. In the above examples, we might consider the information retrieval tools to be a micro level experience, the WWW to be

the middle level and the familiarity with computers to be the macro level experience.

“Most surveys conducted prefer to concentrate only on the middle level of experience, which may not always consist of an adequate set of variables” (Martzoukou, K. 2005). The surveys should have been carried out at the micro level since the users would have acquired common experiences. As described by Martzoukou (2005), “users already have got computer experience and skills to access the Web”. Thus, when studying Web information with a focus on a Web information retrieval system, the experience of micro level might be more appropriate. Hence, this survey was more based on studying users at micro level.

3.2 Sample Population

The survey was administered to staff of Non-IT companies such as ECS Secretaries, Transmetal Ltd, RMB Structured Insurance PCC Ltd and Solutions Knitted for Business Limited. The Non-IT companies were chosen since the management were willing to participate in this survey. One more reason for choosing these non-IT companies is that the staff of these organisations used the Web and computers in their daily processing. Moreover, some IT companies in Mauritius were chosen namely, Ireland Blyth Ltd, Infosys, SITS and Accenture. These companies were chosen since they were known companies in the IT sector in the Mauritian Island. Moreover, the staffs from the above mentioned companies were chosen randomly but the approval of the management was sought before.

3.3 Findings of this survey

3.3.1 Demographics

100 questionnaires were sent with a 77% response.

Age (x):	
18<=x<20	11.7%
20<=x<30	50.7%
30<=x<40	28.6%
40<=x<50	9.0%

Table 3. Population Age

Employment	
IT Background	48.0%
Non-IT Background	41.6%
Unemployed	10.4%

Table 4. Population Employment Background

Internet Knowledge	
Less than 1 year	2.6%
1-2 years	7.8%
3-5 years	26.0%
More than 5 years	63.6%

Table 5. Population Internet Knowledge

3.3.2 Discussion of responses

In response to the question

1. Suppose you need to look for specific information on the Web e.g, 'books written by Enid Blyton'. Where would you look first?

Figure 3 shows the distribution of results. 71.4% of the users responded that they would prefer to use search engines and 19.5% gave interest to meta search engines. However, Web directories had a response of only 6.5% while 2.6% browsed the Web. None used other means than that specified.

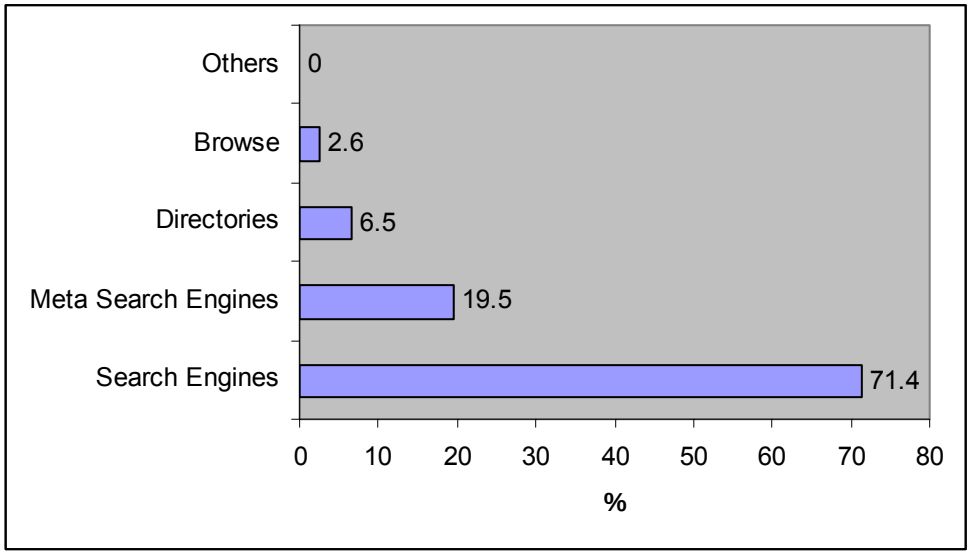


Figure 3. Users' preference of search tools for looking of specific information

This result confirms the habits of users when using search engines in looking for information. As explained by Fallows (2005), from about 68 million American who go online, over 38 million of them will use a search engine. This is a percentage of 55.9%. She also described that people do not always turn to search engines to find answers or information, but they take different paths e.g., going straight to favourite familiar specialised portals.

In response to the question

2. Suppose you have a particular interest and need to keep yourself updated about that particular field. Where would you look first?

We could find that the subscription to mailing lists and use of search engines do not have much difference. As described in Figure 4 the percentage for mailing lists subscriptions was 32.5% while the search engines use was 28.6%. Browsing of the Web was indicated by 22.1% of users and newsgroup 15.6%. However, there was one user (1.2%) who specified the other option.

The results obtained seem to indicate that at least some people subscribe to mailing lists and newsgroups to get information and thus this can help in optimizing the bandwidth usage.

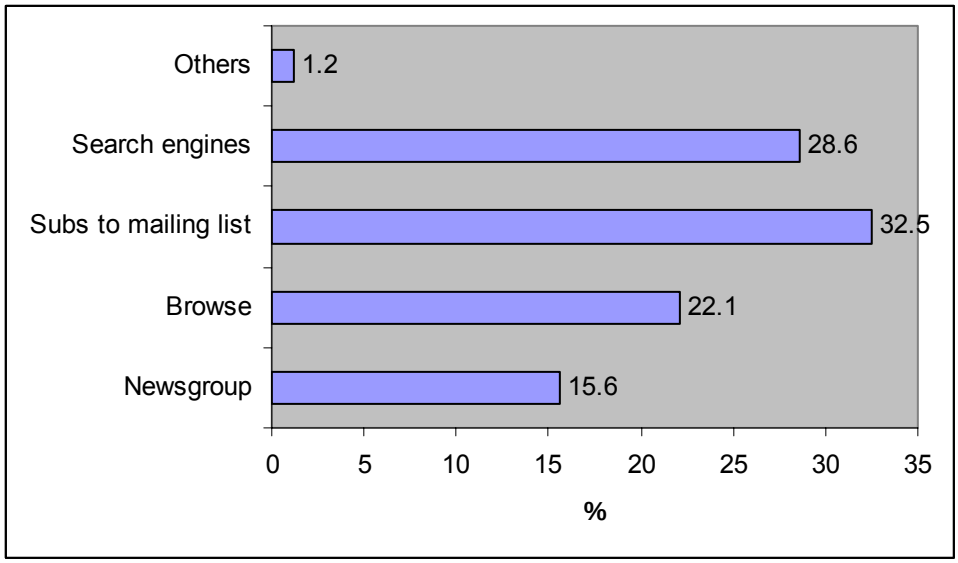


Figure 4. Users' interest of search tools to keep them updated

In response to the question

3. Have you ever used the advanced search option of any search engine?

These results obtained are what were hypothesized from the start of this study. Both the answers were around the same level. Referring to Figure 5, 48.1% gave an affirmative answer whereas 51.9% had a negative one. Users

should be informed about the advantages of the Advanced Search since this helps them to better use the search tool and consequently avoid unnecessary flow of traffic.

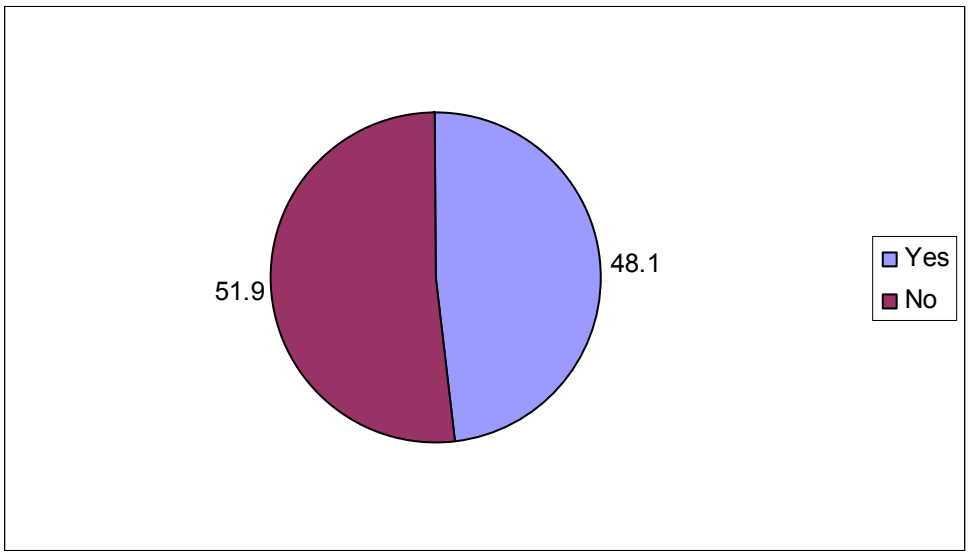


Figure 5. Use of advanced search options

In response to the question

4. When using search engines, how deep do you go in looking for the particular information you are looking for? Referring to your answer, do you find it useful to go deeper than the 1st page while searching?

From the answers to this question, we may conclude that the interest of Web searchers decrease while going into deeper pages of the search result display. From the results, people looking at only the first page constitute 48.1%, 28.6% for the second page, 13.0% for the 3rd/4th page and 10.3% for the 5th page and higher. A further analysis of the answers for the second part of the question showed that 53.3% of users concluded that going deeper than the first page is useless while 46.7% gave an affirmative answer. This has been described in Figure 6.

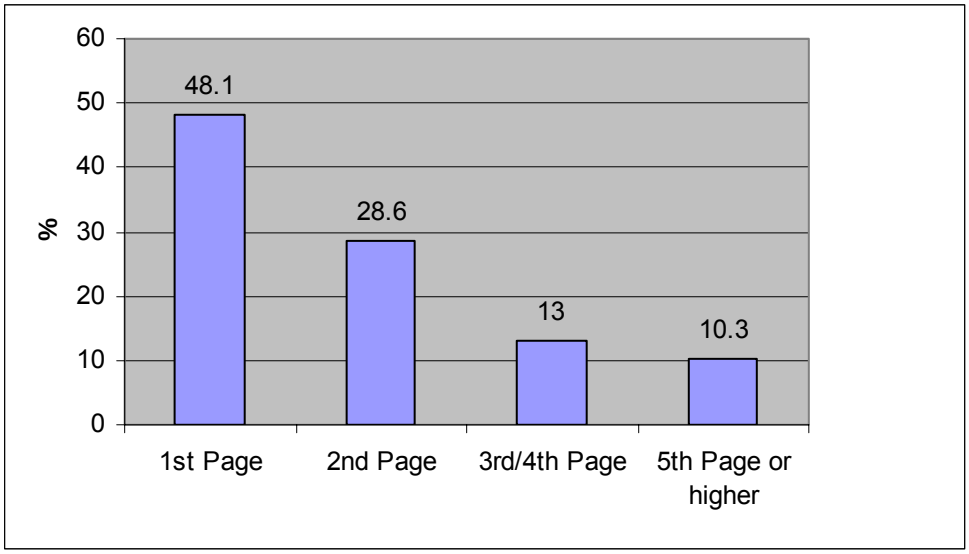


Figure 6. Users' habits of browsing results pages

The same behaviour of Web users was also concluded in the research done by Cacheda and Vina (2000) by examining the log of a Spanish portal. They studied the statistics and found that for 67.88% of the queries performed, only the first screen was checked and in 13.24%, the second page was checked. Hence, taking into account statistics from this survey, it can be said that user behaviour is the same when using a search engine.

In response to the question

5. While browsing, you sometimes come across information that is useful to you but not to your current search. How do you keep these references so that you may use them later?

Figure 7 shows that 55.8% of respondents confirmed that they use bookmarks provided in Web browsers and this is quite encouraging as this shows that users are quite aware about the services provided. However, the result obtained from the memorising option is 33.8 %, and an additional 10.2 % wrote on bits of paper. These two options gave a percentage of 44%, which could be improved.

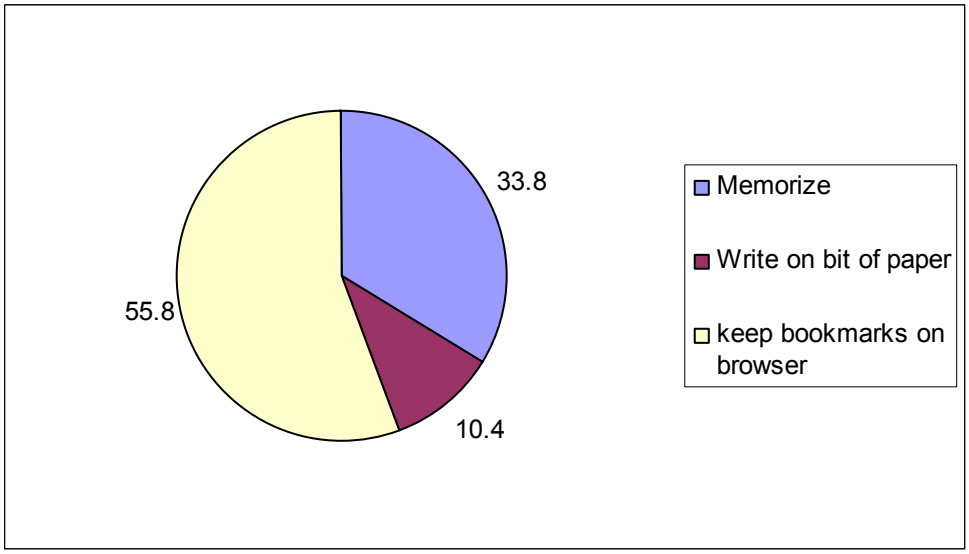


Figure 7. Users' habits of keeping references

In response to the question

6. Nowadays most of the search engines provide toolbars to ease searching for users. Have you ever used any of those toolbars?

As displayed in Figure 8, the response to this question is quite surprising since only 32.5% of the search engine users responded positively. 67.5% said that they never used such a toolbar.

From the respondents who responded affirmatively, notice that most of them preferred using the Google toolbar and Yahoo toolbar, whereas the use of Alta Vista and MSN was quite low.

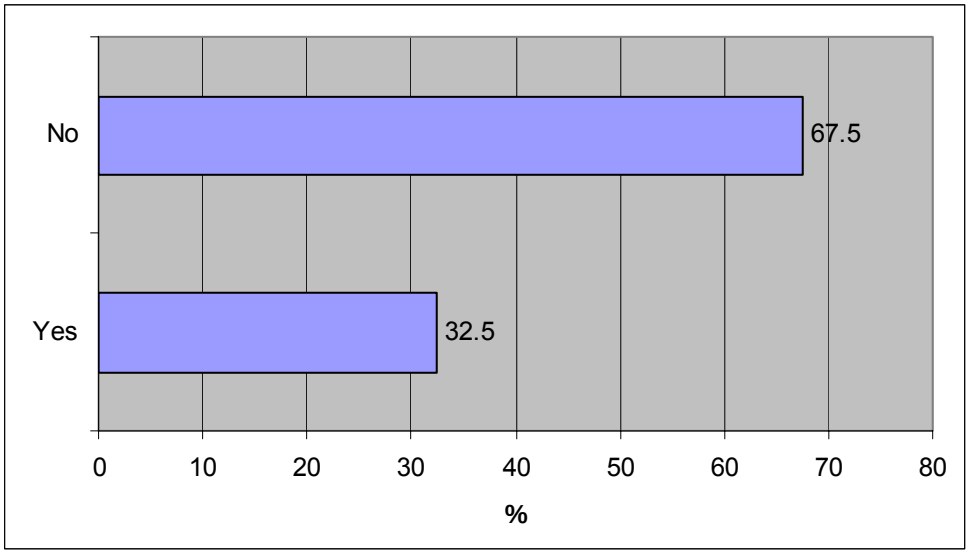


Figure 8. Use of toolbars by search engines users

The results from the survey coincide with that of the survey carried out by iProspect (2004). Their results were that 50.7% of the respondents gave a negative answer to the use of such toolbars.

In response to the question

7. While browsing the Web, you sometimes get unwanted information that is displayed on your screen. Did you know that some toolbars includes a utility that prevents those unwanted screens from popping-up?

The response to this question was not favourable for the optimal use of the bandwidth. As shown in figure 9, 84.4 % of the survey respondents did not know that some tools provided the option of blocking pop-ups as compared to 15.6 % who responded affirmatively to the question. Moreover, those 15.6 % have also used the pop-up blocker at least once.

Information tools do not provide enough information to the user, therefore this is the probable reason why most users are naïve concerning this particular feature.

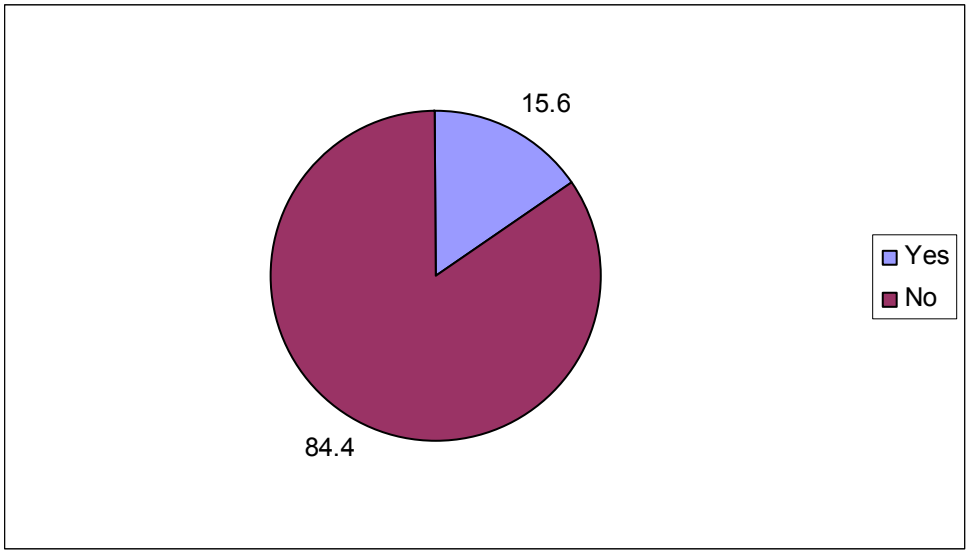


Figure 9. Knowledge about pop-up utility in toolbars

In response to the question

8. Boolean searching is a method of searching the Web by combining words or concepts together. These combinations of words are possible due to the existence of Boolean operators. Have you ever used any of these Boolean operators in your searches? (Examples of Boolean operators are AND, OR, NOT and so on)

The results to this question differ by only 0.8 % (i.e., 1 user from the 77 surveyed). The 'yes' answer resulted in 50.6% and the negative response was 49.4 %. Boolean search can be really important to get useful information. But, as we can notice, nearly 50 % of the respondents never used Boolean operators. Thus, searchers need to be made aware and taught about Boolean operators which need to be clearly displayed on the search page. The results have been displayed in figure 10.

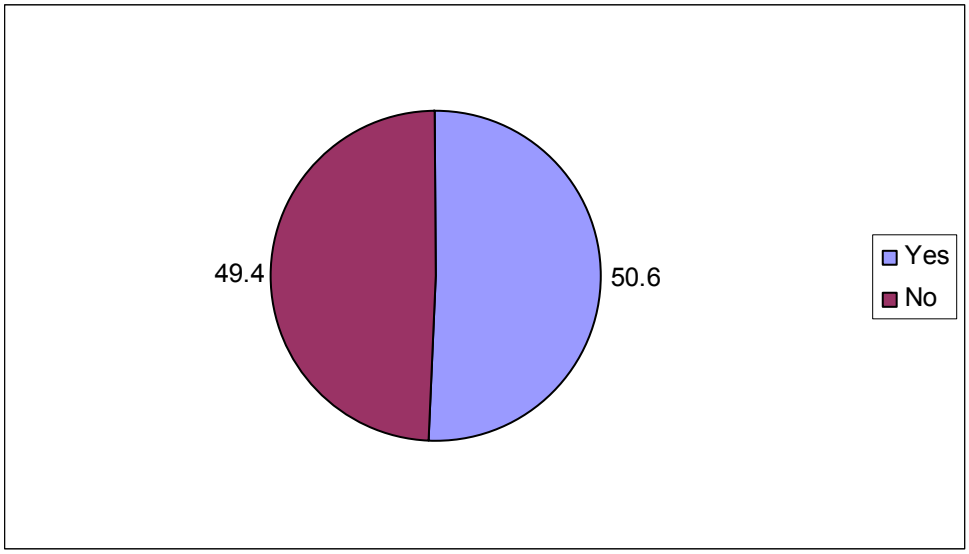


Figure 10. Use of Boolean operators by Web searchers

In response to the question

9. There exists some search engines that can use special characters e.g. * and $ for optimising searches. Are you aware of such specificities?

Figure 11 shows that 18.2 % respondents gave 'yes' as answer to this question whereas the rest i.e., 81.8 % replied negatively. This response is normal from the Web searchers since this information is rarely displayed on the search engines' home pages and, consequently, the users are not aware of such things.

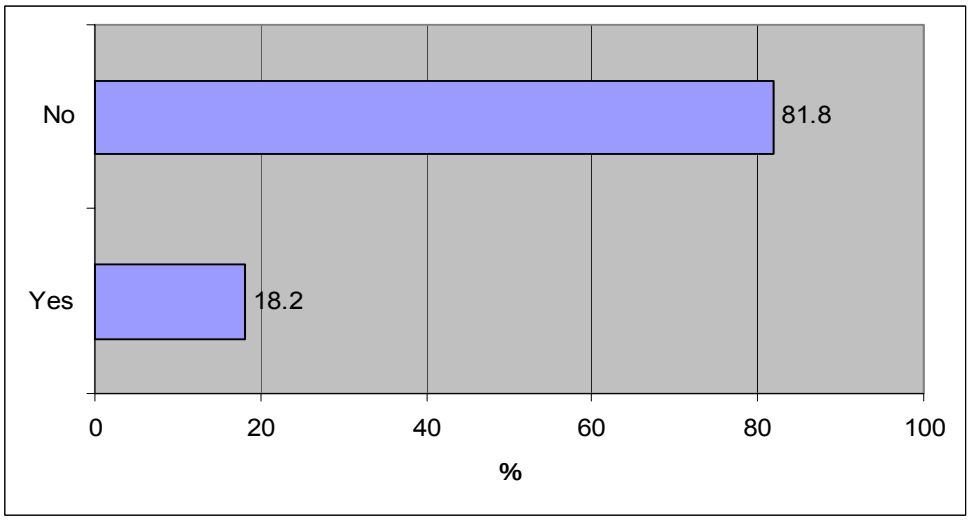


Figure 11. Use of special characters in search engines

In response to the question

10. Search engines are generally designed differently. There exists some search engines which are case sensitive. Are you aware of this?

Question 9 and question 10 are somewhat related since they both depend upon specificities of different systems. Due to the similarity of these questions, we need to see whether the answers also are related. Here, 10.4 % answered no and 89.6 % gave a positive response. Comparing the figures from both questions, we can deduce that they are directly related to each other and thus this confirms that there is a lack of information flow. These issues need to be solved to prevent users from making unnecessary round trips and congesting the network.

In response to the question

11. Most search engines today provide local mirror sites of their search engines. When using search engines, do you take into consideration the use of the mirror sites?

As displayed in figure 12, out of the 77 users who responded to the survey, 83.1% claimed that they do not take into consideration the use of mirror sites while 16.9 % do. We can see that most of the users were unaware of the local sites of search engines and thus, since we have limited bandwidth, the performance of retrieving information from the site is very poor. Reducing the transatlantic traffic would improve the overall performance of information retrieval.

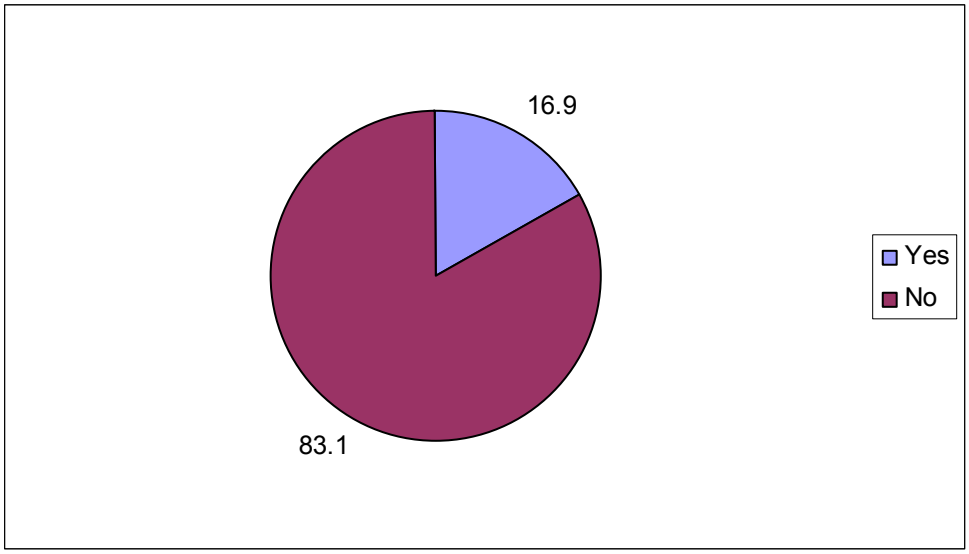


Figure 12. Use of mirror sites of search engines

The answers obtained from this survey concur with the study done by Stobart and Kerridge (1996). They concluded that from 402 respondents they got for their survey, 42% claimed that they did not use a local mirror site of a search engine and 35% were unaware. The remaining 23% confirmed having frequently used such sites.

In response to the question

12. Suppose you look for a specific phrase of the Web, say “Beware the ides of March”. Do you know that you should include that in double quotes “ ” to get the exact phrase in your search?

From the answers obtained, 54.6% gave an affirmative answer while the rest, 45.4%, were negative as shown in Figure 13. Though it might be quite encouraging to see positive feedback for this particular situation, we may conclude that more emphasis should be placed on the proper training of the remaining users in order to enable efficient use of the system.

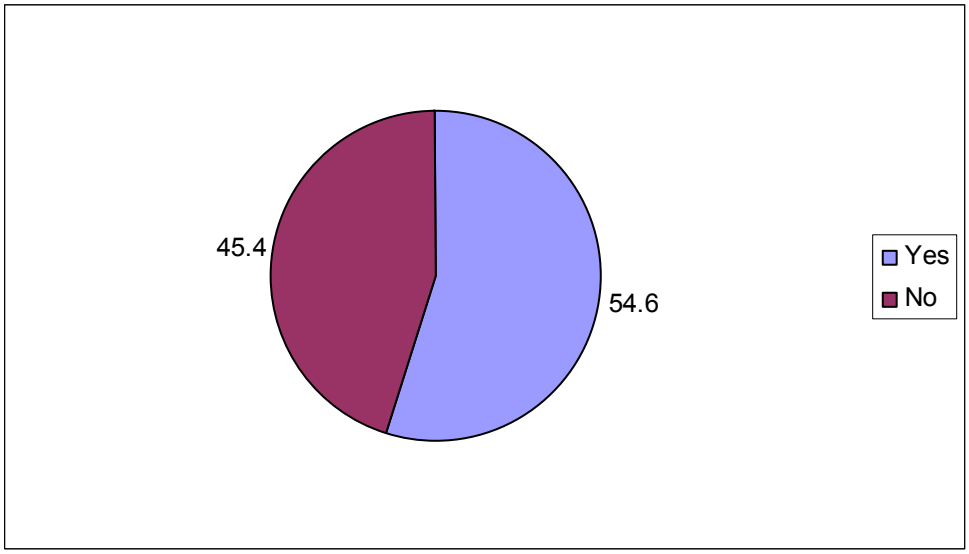


Figure 13. Use of double quotes " " in search engines

In response to the question

13. Do you think that the processing of search engines is done in a reasonable time?

We find from the results obtained that the processing of search engines is rather slow as 66.2% of Web searchers replied negatively while the remaining 33.8% gave an affirmative answer. It is clear from the results that the highest number of users who gave a negative answer (61%) have dial up connections of 56 Kbps. On the contrary, the majority who answered affirmatively to the above mentioned question (15.6%) had a connection of 64/128 Kbps. This has been graphically displayed in Figure 14. The results are simplified in the table:

	512Kbps	64/128Kbps	56Kbps	Unstated	Total
Yes	2.6%	15.6%	7.8%	7.8%	33.8%
No	-	5.2%	61.0%	-	66.2%
				Total	**100%**

Table 6. Results about performance of search engines in different bandwidth environments

Based on the input from the users surveyed, we see that the perceived speed of processing is really slow. To improve the efficiency of users querying information from the Web, either the speed or the user interaction need to be improved. However, due to limited bandwidth, improvements in user interaction may be practical.

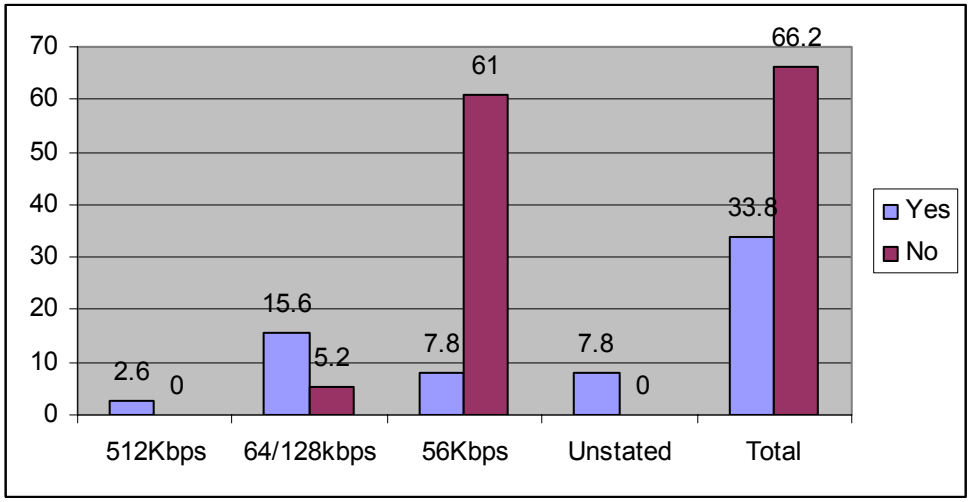


Figure 14. Performance of search engines processing

In response to the question

14. Most of the search engines display only a limited number of results per page. However there exist search engines such as Google which let you specify the number of results. Do you think that all search engines should give you the option to let you specify the number of results per page?

As shown in Figure 15, 81.8% of the Web users said they would like to be given more results per page. However 18.2% did not find it important for the search engines to allow them to display the number of results according to their choice. Moreover, of those who responded affirmatively, their choice was asked concerning the number of results they would like to be displayed. They are summarized in the table below:

No of Results (x)	X<10	10<x<=20	20<x<=50	50<x<=100	x>100	Unstated
Percentage	11.7%	20.8%	36.3%	22.1%	0	9.1%

Table 7. Results about users' preference for display of results on a single page.

Following the survey conducted by Datadial Ltd[4], out of 24 users that were interviewed, if results were not obtained, only 5 went to the 2nd page and the others did reformulate their query. Hence, providing the above option may have been useful.

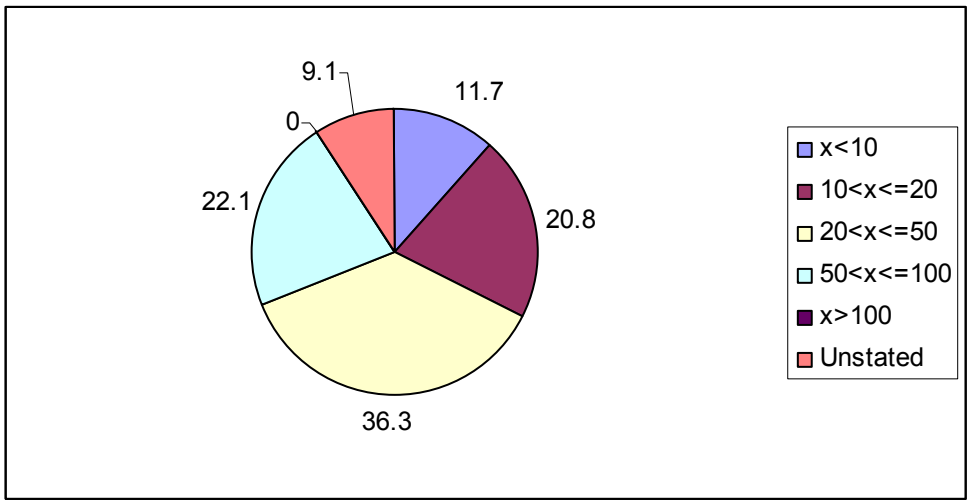


Figure 15. Users preference for display of results on a single page

In response to the question

15. There exists some search engines that provide you with the 'find similar pages option'. Do you think that this option really helps you in your search?

We find that 59.7% of the Web searchers surveyed are unaware about "find similar options" found in search engines and 7.8% said that they have never used such options. The answer is really discouraging since it does not help in optimizing information retrieval. Users need to be better educated in order to allow them to better use the search engines with these options. From the survey conducted by Jansen et al. (1998), on Excite search engines, only 5% of user queries resulted from using the "More Like This" option and it confirms the tendency that users are not aware of the advantages of such features. Increasing the awareness of such features could be useful to the users.

[4] www.datadial.net

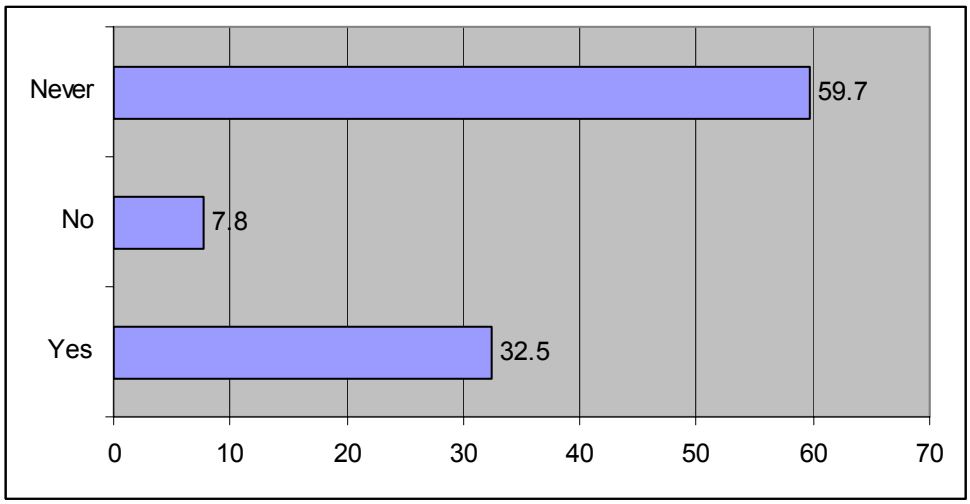


Figure 16. Users' response about using the 'find similar pages option'

In response to the question

16. List the different search engines you use and please specify the reason for choosing the search engines listed.

For the above question, the answers were not well defined since they were in free text format. However, from what we have compiled, most of the users always used the Google and Yahoo sites. Out of the 77 responses obtained, 72 said that they use Google search engine and 69 used the Yahoo search engine. The reasons given for the use of the two search engines are that they are popular, the information retrieval is quicker than other search engines, the homepage is user friendly or they give the most relevant results. Moreover, one of the users interviewed said that they use the Yahoo search engine for picture retrieval.

Other search engines also were given as a choice for the users, among which are MSN (5), AltaVista (7), AskJeeves (2), Teoma(1), Lycos(2) and www.search.com (1). But each of these users had either Yahoo search engine or Google search engine as an option as well.

Hence, from these we can conclude that Web searchers from Mauritius are much more familiar with the Google and the Yahoo search engines and thus advice needs to be provided to them about the other search engines.

In response to the question

17. Do you think that there exist ways for improving the speed of information retrieval from the Web?

From the results compiled, 54.5 % gave a 'no idea' as answer or simply did not answer the question. As explained in the previous question, as the answer was free text, the user did not find it necessary to respond to that particular question. However, 15.6 % responded negatively to the question and the rest, 29.9 %, gave an affirmative answer. There were also many suggestions given about the speed of information retrieval, among which are:

- Use of a powerful computer with good processor and memory since the processing of the information might be done more efficiently on the client side.
- Search engines need to display tips on how to use their search engines.
- Increase of bandwidth to get better connection speed.
- Stop the use of words such as for, of, the, etc.…
- Use of local mirror sites.
- Using of cache (cookies) to keep track of user habits.

All the above reasons given for the optimization of information retrieval concur with this study and search engines should consider all the above to improve usage and minimize bandwidth waste.

In response to the question

18. Meta search engines can be said to be a combination of different search engines into a single application, thus making the results screen more specific. Have you ever used any of such meta search engines?

Figure 17 shows that the response was not favourable since 68.8% of the users surveyed never used such applications. However, the 31.2% who responded affirmatively to this question were asked about their preferences to use meta search engines and a summary of those is:

- Meta search engines avoid Web searchers being flooded with unwanted information.
- Meta search engines provide better link ranking.
- Meta search engines were better than search engines before but nowadays the results returned are almost the same but due to my habit, I still use them.

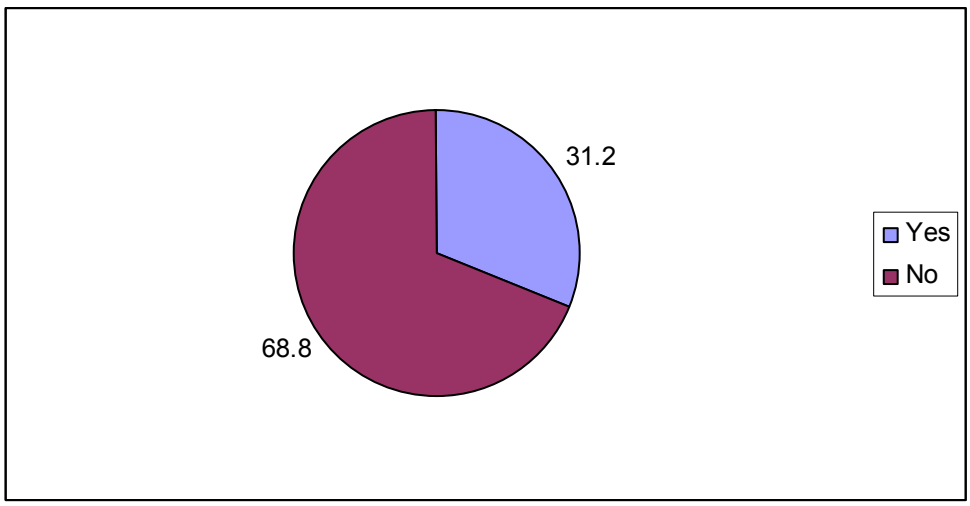


Figure 17. Use of Meta search engines by Web searchers

In response to the question

19. Retrieving information from the Web can be achieved through other ways than search engines. One might be the mailing list. It is a list of email addresses which has been assigned a name. When a message is sent to the mailing list name, it is automatically forwarded to all the addresses in the list. Generally, it gives you useful information to increase your knowledge. Have you ever subscribed to any mailing lists?

Results compiled from this question shows that almost half of the Web users are aware of mailing lists since 51.9% of them claimed that they used these lists whereas the remaining 48.1% never used such facilities as shown in figure 18. However, the use of mailing lists has always been encouraged since it is reasonably efficient in low-bandwidth environments. We might consider the case of TEK (Times Equal Knowledge) search engine developed by Massachusetts Institute of Technology (MIT) [5] who use the same logic of the mailing lists to provide information in poor countries.

However, from the survey carried out, most of the Web searchers did not find the idea of the use of mailing lists to be a good one. The reasons given by them are numerous and summarised below:

- Information is limited in mailing lists.
- Irrelevant information in terms of the users' interest is provided.
- The information displayed in the mail is not well organised in terms of display.
- It might be a good tool for knowledge improvement but it does not clarify the users' search interest most of the time.
- Mails are considered as junk by most users and are most often deleted without even a glance.
- Mailing list is more like an advertisement for websites hosting their information.

[5] www.news.bbc.co.uk/2/low/technology/3065063.stm

Those few who considered the mailing list to be a useful tool gave the following arguments:

- Solutions are provided free without any additional costs.
- Though the information retrieved is sometimes irrelevant, it enriches the Web users' knowledge.

Hence from what has been indicated above, it may be concluded that even though the mailing list is not accepted in most Web searchers' environment, it remains nevertheless a possibly useful tool in low bandwidth situations.

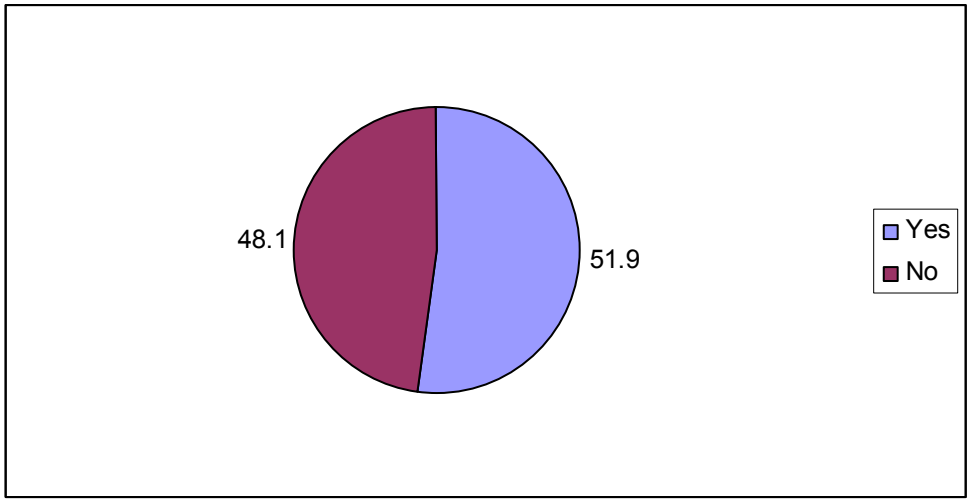


Figure 18. Users' subscriptions to mailing lists

In response to the question

20. Newsgroups work in the same way as mailing lists. But the only difference is that access should be granted to those newsgroups. It is a discussion group that is based on postings about a particular topic. Generally users subscribe to those groups and get information about their field of interest. Did you ever happen to get access to one of those newsgroups?

As explained in Question 19, mailing lists and newsgroups are both considered to be applications using low bandwidth and their use should always be encouraged. However, in question 19, though the response was rather encouraging, we do find it drastic for the use of newsgroups. When the users were interviewed, they were asked if ever they got access to any

newsgroup and only 31.2% confirmed that. The remaining 68.8% never had such access. Newsgroup owners and information seekers could be better informed in order to encourage the use of such tools. Figure 19 illustrates the results.

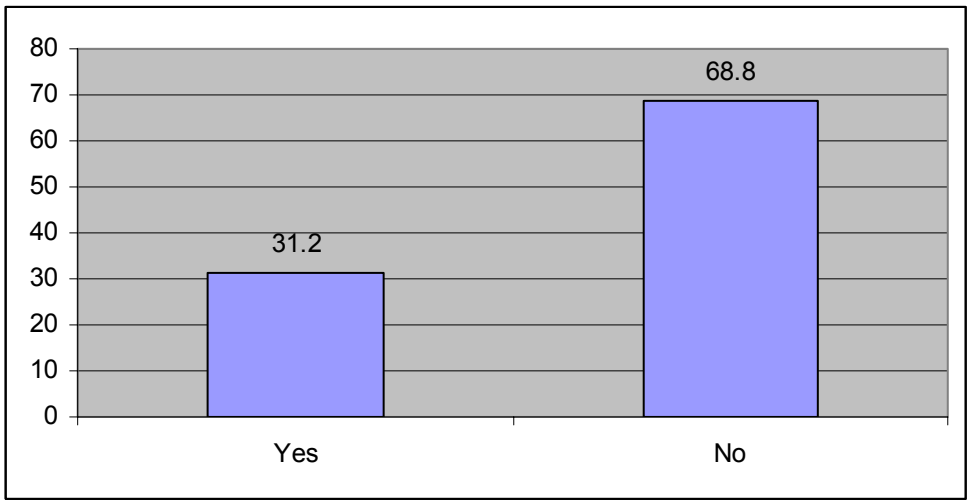


Figure 19. Users' subscriptions to newsgroups

Moreover, the Web users were asked about details of newsgroups they accessed and they were: Delphi newsgroup ((www.info.borland.com/newgroups), Google Newsgroups (www.groups.google.com), www.newgroups.com, and www.cyberfinder.com.

In response to the question

21. With reference to Appendix 1, please give your personal views about the search engines.

The response to this question was quite low i.e., 9.1% of the users were not aware of what type of information should be given. They somehow tried to respond to this question by giving answers and a summary of those are as follows:

(i) How do you find the display on the screen of the search engines?

- The display is good and well planned.

- All search engines have the same display i.e., a search box and a search button.
- Google, Alta Vista and Ask Jeeves search page is more user friendly than the others as it does not display futile information.

(ii) Do you think that there should be any change in the display of these search engines' homepages?
All the users who responded to this question gave a negative answer.

(iii) Is there any similarity or difference between the search engines?
- All the search pages have a search box and a search button.
- The display of the search engines is different but they all perform the same tasks.

3.4 Summary of Findings

For concluding with the evaluation of this survey, we might say that users are tending towards greater simplicity and consistency. Getting successful results from a search tool depends on the knowledge of the user and this is why proper training should be provided to Web search tools' users. As it has been stated by Marchionini (1995), "users' familiarity with search tools depends much on their cognitive abilities". Hence, factors that influence the cognitive abilities of a user for example, training, need to be given attention.

We all know that human beings are careless and due to this characteristic, they make inadequate use of options provided in search tools. Their only concerns when using a search tool is to key in the words in the search box, without even knowing about the specifities of the search tools.

Hence, based on these findings from this survey, we tried to educate and direct users about the use of search tools by making use of a focussed jump page.

CHAPTER 4

Bandwidth Optimisation Search Tool

4.1 Overview

Search engines provide users with a vast amount of information which helps in the development of society. However most search engines today are meant to be commercial and cater primarily for countries with high bandwidth. Hence, for countries like Mauritius, with low bandwidth, specific solutions need to be looked for.

Search engines are developed using complex algorithms in the backend and usually displayed with a textbox and a submit button in the front end. Users thus find it reasonable to input keywords in the textbox to search for information they need to know. They are all unaware of bandwidth and technology that is used for the display of results on their computer screens. Fallows (2005), in her study confirms this by saying that people know little about how engines operate, or about the financial tensions that play into how search engines perform their searches and how they present their results.

Hence, in countries like Mauritius, the bandwidth use needs to be optimised either from the backend or the front-end of the search engine. However, since the algorithms and source codes remain in the possession of the search engines developers, the bandwidth needs to be optimized from the user side. The user interaction is proportional to the performance of a search tool, hence the bandwidth use. This has been confirmed Zhang, X., Li, Y., and Jewell, S. (2005) in their paper where they described that search knowledge reflects the degree a user knows how to plan his/her searching and the search knowledge is an important factor for a successful search.

In Mauritius and countries with simpler network architecture, the display of search engines like Google or Yahoo needs to be worked upon so that users find it easy to use in order to prevent unnecessary clicks. A small tool was thus developed using HTML and JavaScript in order to allow users to retrieve

information with better access and information on search engine options. It is available at www.geocities.com/aballuck.

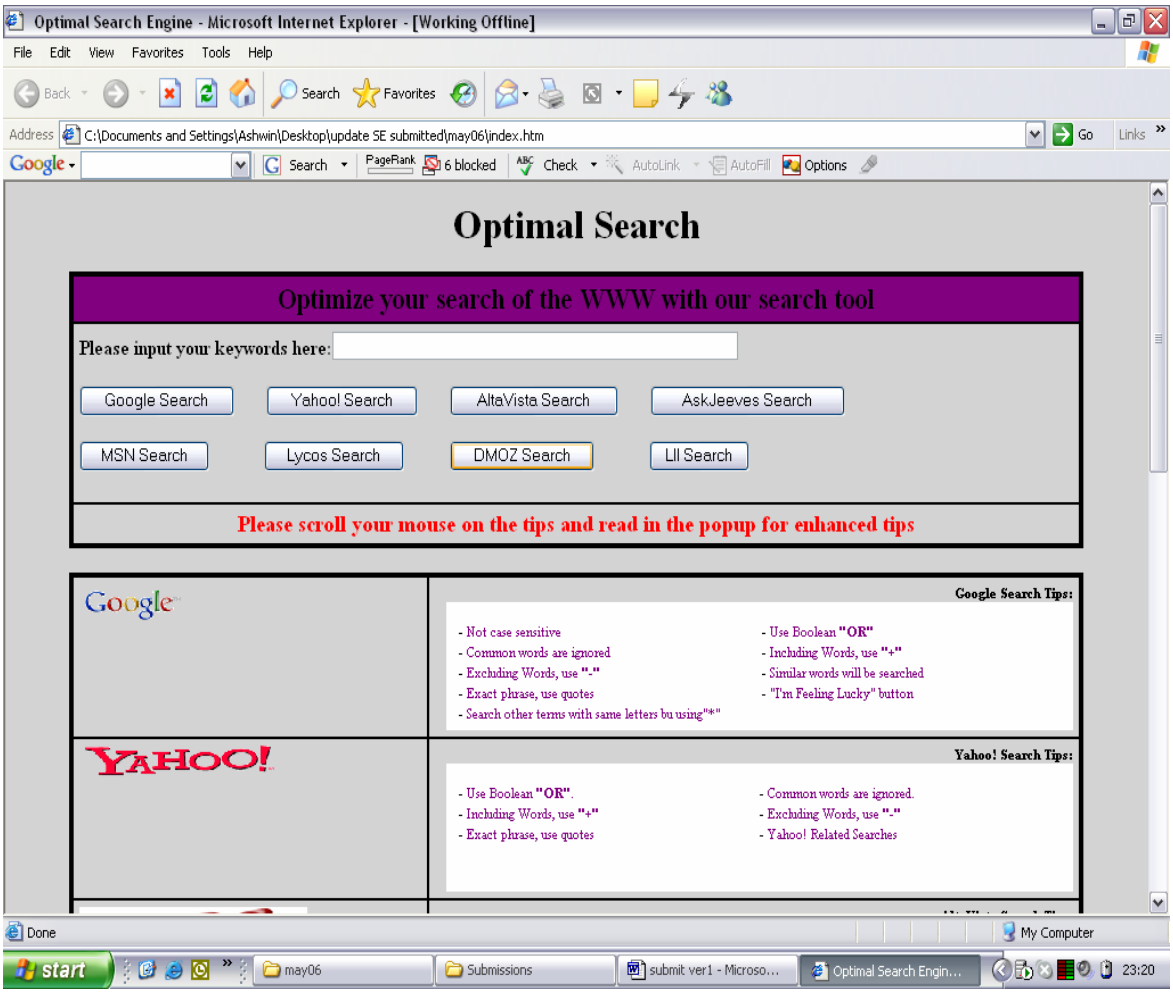


Figure 20. Jump Page developed to assist users

The display was designed in order to allow users to quickly and efficiently retrieve information with fewer clicks. Different well known search engines were chosen to be included on this jump page, among which are Google, Yahoo!, AltaVista, MSN, AskJeeves and Lycos. Moreover, two well-known Web directories were included in this page namely, DMOZ and Librarians' Internet index. We have also included some quick search tips that will generally help the Web users to formulate their queries.

4.2 Further evaluation

Having carried out an initial survey and consequently developed a jump page for search engines, further evaluation of users is important in order to assess the effectiveness of new tools. As Martzoukou described in his study, “the most important part of studying user’s behaviour is through observation”. “This method can reveal information that is not easily discovered by other techniques” (Martzoukou, K. 2005). Hence, I suggest that to better assess any changes, users should be observed and later questioned.

For this evaluation, I have chosen a sample of 10 users, of which 5 have an IT background and the remaining 5 are from a non-IT field. A brief description of the jump page - about the graphical user interface and the different search engines that have been included on the page - was given to the users. The different specificities of each individual search engine were explained and that the tips provided would give them a brief overview of how to use the search tools. A briefing on the keys to successful searching provided in the Spider’s Apprentice[6] was also done as follows:

- All the users were briefed about where to look for information i.e., search engines or directories. They were also told about the difference between them and which engine of the jump page is more like a search engine or a directory.
- They were asked to fine-tune their keywords when formulating their queries.
- The query by example was also explained, for example, the use of ‘find similar sites’ on the results page.
- The answers need to be anticipated i.e., how the most useful page would look like was indicated to them.

After such an exercise, the users were asked to search for any information from the Web using the jump page and they were observed in carrying out

[6] www.monash.com/spidap.html

their tasks. The first observation was that the users were careless and restless and clicked on any of the search buttons without knowing the specific details of that engine. The earlier explanations were repeated and this was fruitful since there was an improvement in their way of processing.

Moreover, the search carried out by the 5 IT persons was more successful than the other 5. They knew how to formulate queries and obtained their results in less time than the 5 non-IT persons. However, the former were also successful in their search but they needed to go through the tips provided first and later reformulate their queries. The tips have helped them to learn about searching and thereby they used less of a trial-and-error approach to the search activities. Though the time taken was more for the non-IT background users, the number of clicks by all the users was somewhat the same. Thus, as we described before, the bandwidth congestion depends on the number of clicks rather than the time taken to click on a link. Hence, since the number of clicks is the same for both classes of users, it implies that the bandwidth use is the same.

It was also noticed that most of the users used the Google engine for their search. There are two reasons which explain this. First, since the Google button is in the first position, they might have clicked the button more frequently. Second, according to the users, they considered Google to be the most famous search engine and this is why this engine was more used that the others.

By observing them using the jump page, a better understanding of how the users used this tool was obtained. Moreover, after such observation, the users were asked some questions and asked for their views on the tool.

1. Have you used any of the syntax in your search?

Yes – 10 users No – none

They all claimed to have used the syntax and it proved to be useful. However, they said that they needed more practice in the use of the Boolean operators.

2. Do you prefer all search engines on a single page?

Yes – 9 users No – 1 user

We can see that there is only one user who said that he does not prefer all search engines on a single page. The reason given is that he does not find it useful since he always uses Google to retrieve information from the Web. However, those whose answers were affirmative gave multiple reasons for their choice. They said that it is better since multi tasking is possible. Also, since all search engines are on a single page, it saves time by eliminating swapping among different search engines. One more reason for such choice is that they may compare between different search engines and know which of them yields better results for certain searches.

3. Do you prefer all information and tips displayed on the front page?

Yes – 8 users No – 2 users

From this question also, we see that the majority of the users gave an affirmative answer and they gave several reasons for their decisions e.g., ”The information on the front page helps us since we need not browse additional pages to get required information, hence saving time”. One of the non-IT users indicated that the tips provided helps him since he is not familiar with search engines. However, one of the users who gave a negative answer said that all information displayed on a single page does not help since, when looking for information, he does not have much time for reading tips and information and thus retrieves information by trial and error.

The IT persons evaluated were really conversant with the jump page and one of the reasons that can support this behaviour is the knowledge acquired in their professional background. However, the remaining 5 users needed some

coaching on how to query in their search. To conclude with this further evaluation, it can be said that providing information to users about specifities of search tools proved useful. However, as we may all be aware, changing human habits is not easy and thus constant information need to be provided to them to allow efficient use of search tools.

CHAPTER 5

5.1 Conclusion

The purpose of this study was to look into how to optimise the use of search tools in low bandwidth environments which is present in developing countries such as Mauritius. To better assess the bandwidth use by users of search tools in such countries, I tried to evaluate them about their way of looking for information. The response to how users look for information indicated that most of them were not well trained in using search tools. Later I developed a jump page and a sample of users, i.e. IT and non-IT background users were chosen for further evaluation. They were asked to use the search tools developed, which had a new Graphical User Interface. From this further assessment, the result was satisfactory as almost all the users were using the tools as per the expectation of this survey.

This study has derived that users in the developing countries do not well understand the advanced features that are present in the search tools. The knowledge of the users might be improved through training but the other factor that may influence the users' behaviour is the introduction of focused tools. Hence, this will have a direct effect on the bandwidth usage.

To conclude, we can say that although Mauritius is considered to be one among the most developed Nations of Africa, Internet facilities have not kept pace. Hence, to increase the efficiency of retrieving information, the infrastructure needs to be better developed. However, this is improbable in the short term. This study has helped to demonstrate that for immediate benefit, users need to be well informed and tools need to be provided to them so that bandwidth is wisely used.

5.2 Future Work

This study gives a basic overview of how Web searchers are inexperienced in using search tools leading to a waste of precious bandwidth. This was merely the start of a new research direction since no proper study was performed on such topic before. All studies carried out were to optimise the search tools from the back end instead on trying to optimise the searchers' habits.

Hence, the scope for future research in this field is immense. Development in terms of technology was never given a proper start in Africa and it will take years for a boom to happen in this sector. Hence, providing careful assistance to users of the Web might make us use the precious bandwidth optimally in African communities. We thus suggest that this study be extended to the whole African Region with a larger sampling since this study was carried out in the Mauritian environment and the results might not hold true for other regions in Africa.

Moreover, the search tools need to be well defined and planned since their use affect the bandwidth directly. If more planning is performed on the Graphical User Interface to make the features of the search tools visible to the users, it will help to get a better grasp of the basic of the search tools and this will consequently optimise bandwidth use.

BIBLIOGRAPHY

- Allen, B.L. (1991) "Topic knowledge and online catalog search formulation." *Library Quarter, 61(2), 188-213.*

- Ackbarally, N. (2000) "Underwater Fibre-Optic cable installed in Mauritius". *Panafrican News Agency.*

- AfrISPA, (2005) "African Internet Service Provider's Association" AfrISPA, African regulatory index reports, Mauritius report, (region two), June 2005. *[Available at: http://www.catia.ws/Documents/File/MU_Final%20Draft%20Version%20vers%206.0.doc].*

- Altavista – www.altavista.com

- Almeida, R.B. and Almeida V.A.F. (2004) " A community Aware search engine". *WWW2004, May 17–22, 2004, New York, New York, USA. ACM 158113844X/04/0005*

- Ask Jeeves – www.ask.com

- Baeza-Yates, R., Ribeiro-Neto, B. (1999) "Modern information Retrieval". *Addison-Wesley, Reading ISBN 0-201-39829-X Pg 269-271, 384*

- Barbosa, A. M. (2000) "Overview of text summarisation in the context of information retrieval and interpretation: Applications for Web pages summarization". *[Available at: http://www.abarbosa.org/docs/web_pages_summarization.pdf].*

- Bates, M.J. (1998) "Indexing and access for digital libraries and the Internet; Human database and domain factors." *Journal of the American Society for Information Science, 49(13), 1185-1205.*

- Bhavani, S K. (2001) “General and efficient strategies for information Retrieval*”. [Available at: www.personal.si.umich.edu/~bhavnani/papers/Bhavnani_et_al-ASIST-2001-(short_paper).pdf]*

- Cacheda, F. and Vina, A. (2000) “Understanding how people use search engines: A statistical analysis for e-business”. *[Available at www.tic.udc.es/~fidel/docs/publications/e-2001.pdf].*

- Campbell, K. (2000) “Click through the clutter.” Profit. May 2000. Lexis Nexis. September 8, 2000. *[Available at www.web.lexis-nexis.com/universe].*

- Chung, I-Hsin (2000) “Strategies for dealing with high and low bandwidth connections”. *[Avaliable at: www.otal.umd.edu/UUGuide/ihchung].*

- CIA (2007) ‘CIA World Factbook’, Central Intelligence Agency (CIA) *[Avalilable at : https://www.cia.gov/cia/publications/factbook/print/mp.html]*

- ClickZ Experts Advice & Opinions – *www.clickz.com*

- Conducting research on the Internet – *www.library.albany.edu/internet/research .html*

- Datadial Ltd – www.datadial.net *[Available at 181/183 Warwick Road – Warwick House – London W14 8PU. Email to: info@datadial.net]*

- Ding, Y. et al. (2000) “Bibliometric Information Retrieval System (BIRS): A Web search interface using Bibliometric Research results”. *Journal of the American Society for Information Science 51(13): 1190-1204.*

- DMOZ Directory *www.dmoz.org*

- Fallows, D. (2005) "Search engine Users" *PEW Internet & American Life Project. [Avaliable at www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf].*

- Fidel, R., Davies, R.K., Douglas, M.H, Holder, J.K, Hopkins, C.,Kushner, E. J., Miyagishima, B.K. & Toney, C.D. (1999) "A visit to the information mall; Web searching & behaviour of high school students." *Journal of the American Society for Information Science 50, 24-37.*

- Gaines, B.R., Chen, L.L. & Shaw, M.L.G (1997) "Modelling the human factors of scholarly communities supported through the Internet and the www." *Journal of the American Society for Information Science, 48(11), 987-1003.*

- Google – *www.google.com*

- Gordon, M. & Pathak, P. (1999) "Finding information on the www; The retrieval effectiveness of search engines." *Information Process & Management 35, 141-180.*

- Granka, L.A., Joachims, T. and Gay, G. (2004) "Eye-tracking Analysis of User Behaviour in WWW search". *SGIR '04, July 25-29,2004. Copyright 2004 ACM 1-58113-881-4/04/0007.*

- Gray, T. A. (2004) "How to search the Web. A guide to search tools". *[Available at: http://daphne.palomar.edu/TGSEARCH].*

- Gudivada, V., Ragahvan, V., Grosky, W. and Kasanagottu, R. (1997) "Information Retrieval on the world wide Web". *IEEE Internet Computing, 1997, 1(5), 58-68.*

- Information Retrieval: From the traditional way to the Web - L'express newspaper – Mauritius 11th April 2005.

- iProspect (2004) "iProspect search engine User Attitudes" *[Avaliable at: www.iprospect.com].*

- Invisible Web – *www.invisible-Web.com*

- ITU (2004) "The Fifth Pillar: Republic of Mauritius ICT Case Study", International Telecommunications Union (ITU), Place des Nations, Geneva 20, Switzerland

- Jansen, B.J., Spink, A., Bateman, J. & Saracevic,T. (1998) "Real life information retrieval: A study of user queries on the Web." *SIGIR Forum 32, 11 5-17.*

- Jansen, B.J. and Pooch, U. (2000). Web user studies: A review and framework for future work. *Journal of the American Society of Information Science and Technology*. *52, 3, 235-246.*

- Jansen, B.J., Spink, A. & Saracevic,T. (2000) "Real life, real users and real needs. A study and analysis of users' queries on the Web." *Information processing & Management, 36(2), 207-227.*

- LII – *www.lii.org*

- Liu, J. (1998) "Guide to meta search engines." *BF Bulletin Special libraries Association, Business and Finance Division, 107 (Winter 1998), 17-20.*

- Lycos – www.lycos.com

- Marchionini, G.L. (1995) "Information seeking in electronic environments." *Cambridge, Cambridge University Press.*

- Martzoukou, K. (2005) "A review of Web information seeking research: considerations of method and foci of interest" *Information Research*, **10**(2) paper 215. *[Available at http://InformationR.net/ir/10-2/paper215.html].*

- MSN – *www.msn.com*

- Nelson, M. R. (1997) "We have the Information you want, but getting it will cost you: being held hostage by information overload." ACM Crossroads 1(1) *[Available at www.acm.org/crossroads/xrds1-1/mnelson.html].*

- Pejtersen, A.M. & Fidel, R. (1998) "A framework for work-centered evaluation and design; a case study of IR on the Web." *Working paper for Multimedia Information Retrieval applications (Mirs) workshop, Grenoble France.*

- Raux, A (2003) "EUREKA: Dialogue-based Information Retrieval for very low bandwidth environments". *11-743 Advanced IR Seminar Final Report.*

- Savoy, J. (2000) "Information Retrieval on the Web". *[Available at www.unine.ch/info/Gi/papers/SI.pdf].*

- Savoy, J & Picard, J. (2001) "Retrieval effectiveness on the Web." *Information processing & management 37(4), 543-569*

- Schneiderman, B. (1998) – Designing the user-interface: strategies for effective human computer interaction. 2nd Ed. Addison-Wesley Publishing Company Inc., Reading, MA, 1992

- Second Generation searching on the Web – *www.library.albany.edu/internet/second.html*

- Shen, D. et al. (2004) "Web-page classification through summarization". Proc. of the 27th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR' 04 *[Available at http://research.microsoft.com/~zhengc/papers/SIGIR2004_p242-shen.pdf].*

- Stelmaszewska, H. and Blandford, A. (2005) "Patterns of interactions: user behaviour in response to search results". *[Available at www.uclic.ac.uk/annb/docs/stelmaszewska].*

- Stobart, S. and Kerridge,S. (1996) "An investigation into world wide Web search engine use from within the UK – Preliminary findings*". [Available at www.ariadne.ac.uk/issue6/survey]*

- Thies, W. et al. (2002) "Searching the world wide Web in low-connectivity communities". *Laboratory of Computer Science, MIT. TEK Experiment. [Available at www.cag.lcs.miyt.edu/commit/papers/02/www2002-tek.pdf].*

- The Future of search engine optimising: Theme engines by Robin Nobles – www.searchenginesworkshops.com

- The Spider's Apprentice – A helpful guide to Web search engines – *www.monash.com/spidap.html.*

- Van Rijsbergen, C.J. "Information Retrieval" Department of Computer Science, University of Glascow *[Available at www.dcs.gla.ac.uk/keith/Preface.html]*

- Wilson, T.D (2000) "Human Information Behaviour". *Special Issue on Information Science Research Volume 3 No 2, 2000.*

- Woodward, J. (1996) “Cataloguing and classifying information resources on the Internet.” *Annual review of Information Science & technology, 31, 189-219.*

- Wurman, R.S. (1989) “Information Anxiety”. *New York: Doubleday.1989.*

- Wylic, I. (2002) “Fast-company – How smart people work*.” [Available at www.fastcompany.com].*

- Yahoo – *www.yahoo.com*

- Zhang, X. M., Angheslescu H.G.B. & Yuan X. (2005) “ Domain knowledge, search behaviour and search effectiveness of engineering & science students; an exploratory study.” *Information Research 10(2) paper 217. [Available at http://InformationR.net/ir/10-2/paper217.html]*

- Zhang, X., Li, Y., & Jewell, S. (2005) “Design and evaluation of a prototype user interface supporting sharing of search knowledge in information retrieval.” *Proceedings of the 68th Annual Meeting of the American Society for Information Science and Technology.* Information Today, Medford, NJ, pp.

APPENDIX - Questionnaire

Dear Colleague,

Information Retrieval form the Web relies on tools such as search engines and their use by individual information seekers. In low-bandwidth environments, the use of these tools may need to be optimised by a combination of policy and technical solutions and this will be the subject of my research study. This survey is part of the information gathering phase to ascertain what the current habits and needs of the users are. Thus, this will be a prelude to testing directed optimisations.

Please answer all the questions below as accurately as possible. Your answers will be kept strictly confidential.

Thank you.

Regards,
Ashwinkoomarsing BALLUCK
Student Number: BLLASH003
Department of Computer Science
University of Cape Town

Your personal information

Age: ……

Profession: ……

'Internet and Web Search' Knowledge:

□ Less than 1 year

□ 1-2 years

□ 3-5 years

□ More than 5 years

Questions:

1. Suppose you need to look for specific information on the web e.g., 'books written by Enid Blyton'. Where would you look first?

□ Search Engines □ Directories

□ Meta Search Engines □ Browsing the web

□ Other (Please Specify) ……..

2. Suppose you have a particular interest and need to keep yourself updated about that particular field. Where would you look first?

□ Newsgroup □ Subscribe to mailing lists

□ Browsing the Web □ Search Engines

□ Other (Please Specify) ……..

3. Have you ever used the advanced Search option of any search engine?

□ Yes (Please specify which Search engine) …….

□ No

If your answer to above is yes, has it been useful to you?

□ Yes

□ No

4. When using Search Engines, how deep do you go in looking for the particular information you are looking for?

□ 1st page

□ 2nd page

□ 3rd page / 4th page

□ 5th page or higher

Referring to your answer, do you find it useful to go deeper than the 1st page while searching?

□ Yes

□ No

5. While browsing, you sometimes come across information that is useful to you but not to your current search. How do you keep these references so that you may use them later?

□ Memorize

□ Write on bit of paper

□ Keep bookmarks on your browser

6. Nowadays most of the search engines provide toolbars to ease searching for users. Have you ever used any of those toolbars?

□ Yes (Which toolbar(s)?) ……

□ No

7. While browsing the web, you sometimes get unwanted information that is displayed on your screen. Did you know that some toolbars includes a utility that prevent those unwanted screens to pop-up?

□ Yes

□ No

If your answer is yes, have you ever used a pop-up blocker?

□ Yes

□ No

8. Boolean searching is a method of searching the web by combining words or concepts together. These combinations of words are possible due to the existence of Boolean operators. Have you ever used any of these Boolean operators in your searches? (Examples of Boolean operators are AND, OR, NOT and so on)

□ Yes

□ No

9. There exists some search engines that can use special characters e.g. * and $ for optimising searches. Are you aware of such specificities?

□ Yes

□ No

10. Search engines are generally designed differently. There exists some search engines which are case sensitive. Are you aware of this?

□ Yes

□ No

11. Most search engines today provide local mirror sites of their search engines. When using search engines, do you take into consideration the use of the mirror sites?

□ Yes

□ No

12. Suppose you look for a specific phrase of the web say “Beware the ides of March”. Do you know that you should include that in double quotes “ ” to get the exact phrase in your search?

□ Yes

□ No

13. Do you think that the processing of search engines is done in a reasonable time?

□ Yes

□ No

What is your internet connection provided by your ISP? ……...

14. Most of the search engines display only a limited number of results per page. However there exist search engines such as Google which let you specify the number of results. Do you think that all search engines should give you the option to let you specify the number of results per page?

□ Yes

□ No

If your answer is yes, please indicate the number of results that should be displayed. …….

15. There exist some search engines that provides you with the 'find similar pages option'. Do you think that this option really helps you in your search?

□ Yes □ Never used this option

□ No

16. List the different search engines you use and please specify the reason for choosing the search engines listed.

……

……

……..

……..

……..

17. Do you think that there exist ways for improving the speed of information retrieval from the Web?

……..

……..

……..

……..

……..

18. Meta search engines can be said to be a combination of different search engines into a single application, thus making the results screen more specific. Have you ever used any of such Meta search engines?

□ Yes

□ No

If your above answer is affirmative, please give details about your preferences for using search engines or Meta search engines.

……

……..

19. Retrieving information from the Web can be achieved through other ways than search engines. One might be the mailing lists. It is a list of email addresses which has been assigned to a list. When a message is sent to the mailing list name, it is automatically forwarded to all the addresses in the list. Generally, it gives you useful information to increase your knowledge. Have you ever subscribed to any mailing lists?

□ Yes

□ No

Personally, do you think it helps users in getting fruitful information?

……..

……..

20. Newsgroups work in the same way as mailing lists. But the only difference is that you should be allowed to get access to those newsgroups. It is a discussion group that is based on postings about a particular topic. Generally users subscribe to those groups and get information about their field of interest. Did you ever happen to get access to one of those newsgroups?

□ Yes

□ No

If your above answer is affirmative, please give details about such newsgroups.

……

……..

21. Reference to Appendix 1, please give your personal views about the search engines.

i) How do you find the display on the screen of the search engines?

……

……

……

ii) Do you think that there should be any change in the display of these search engines' homepage?

……

……

……

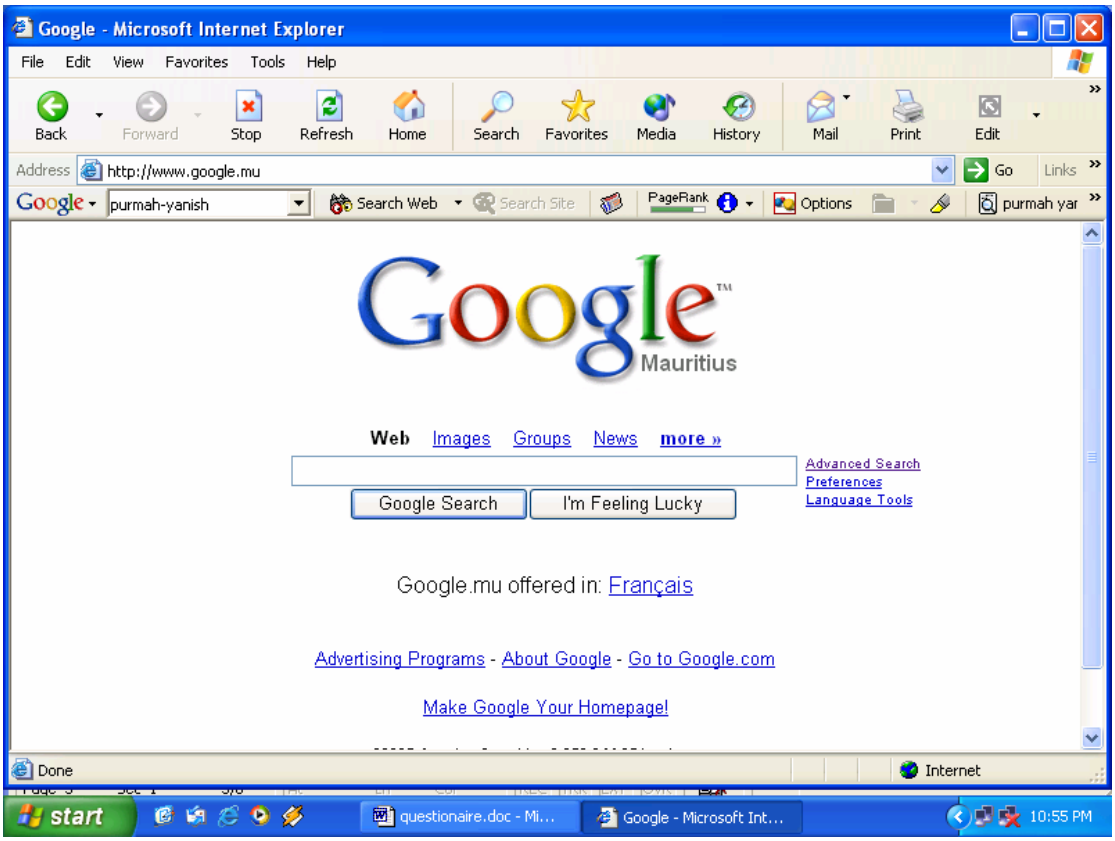
iii) Is there any a similarity or differences between the search engines?

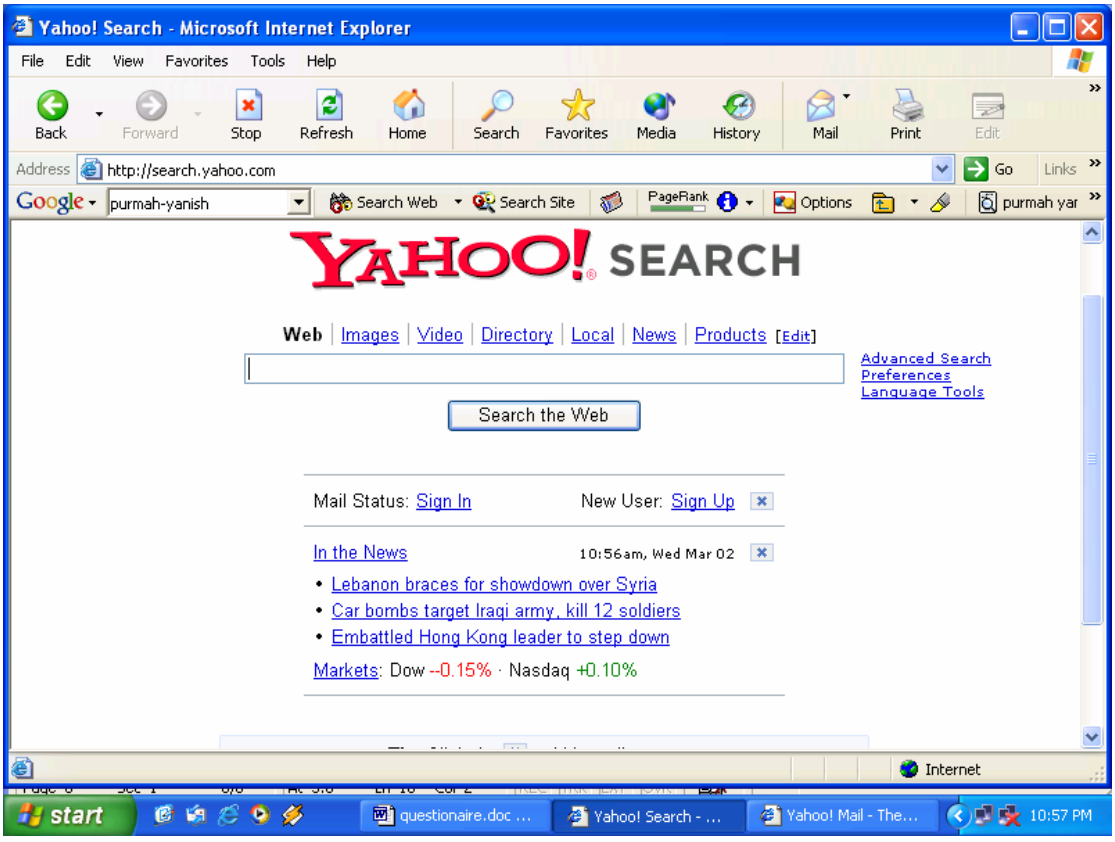
……

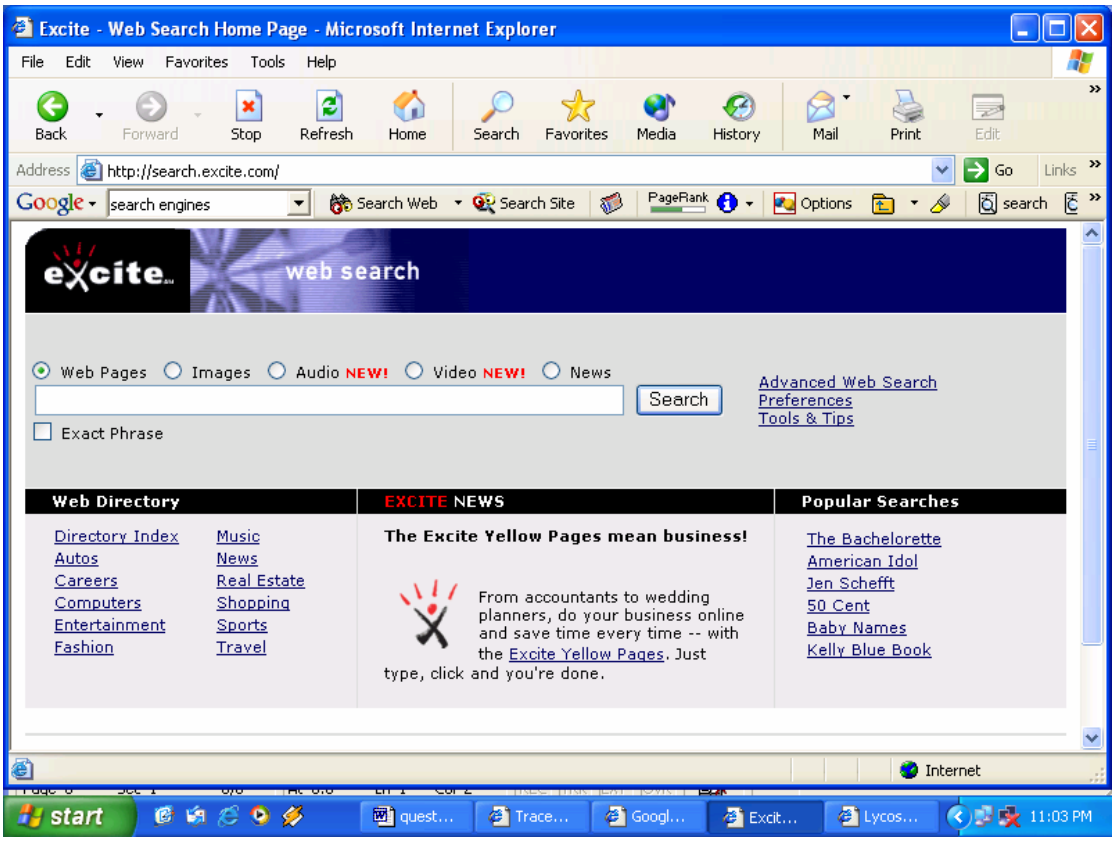
……

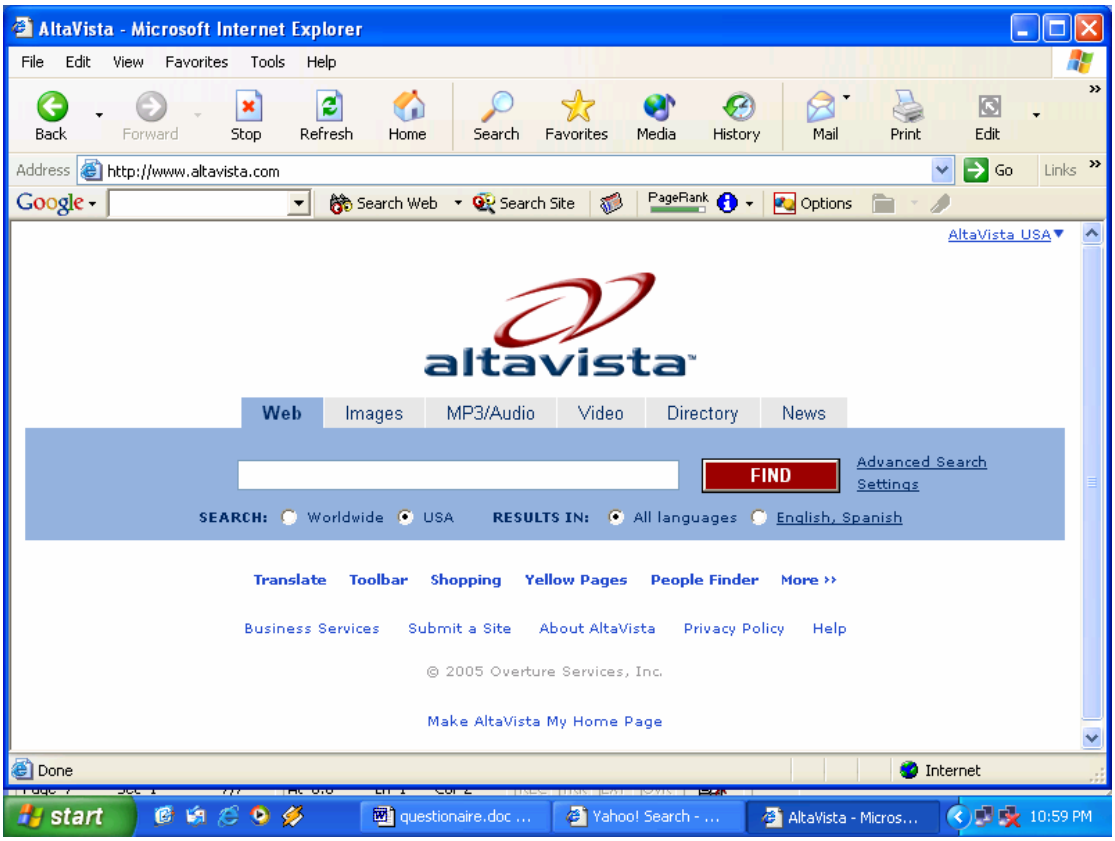
……

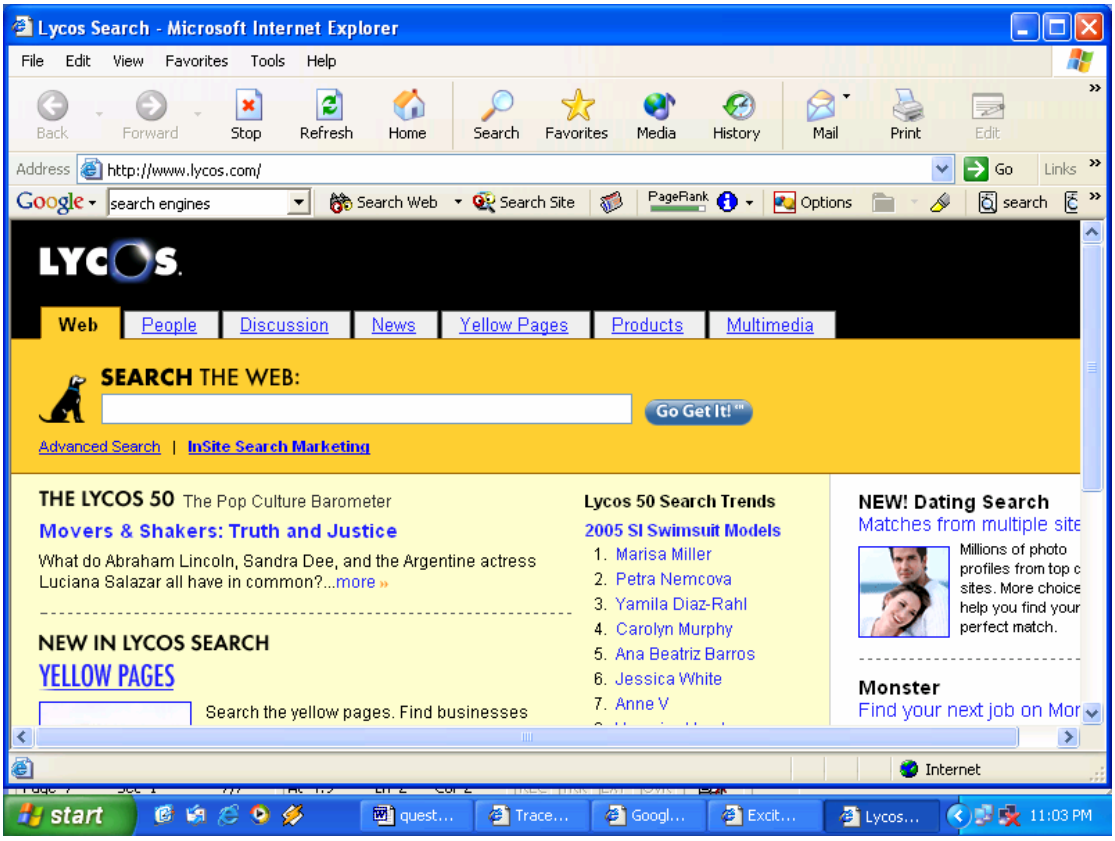
Screens Print:

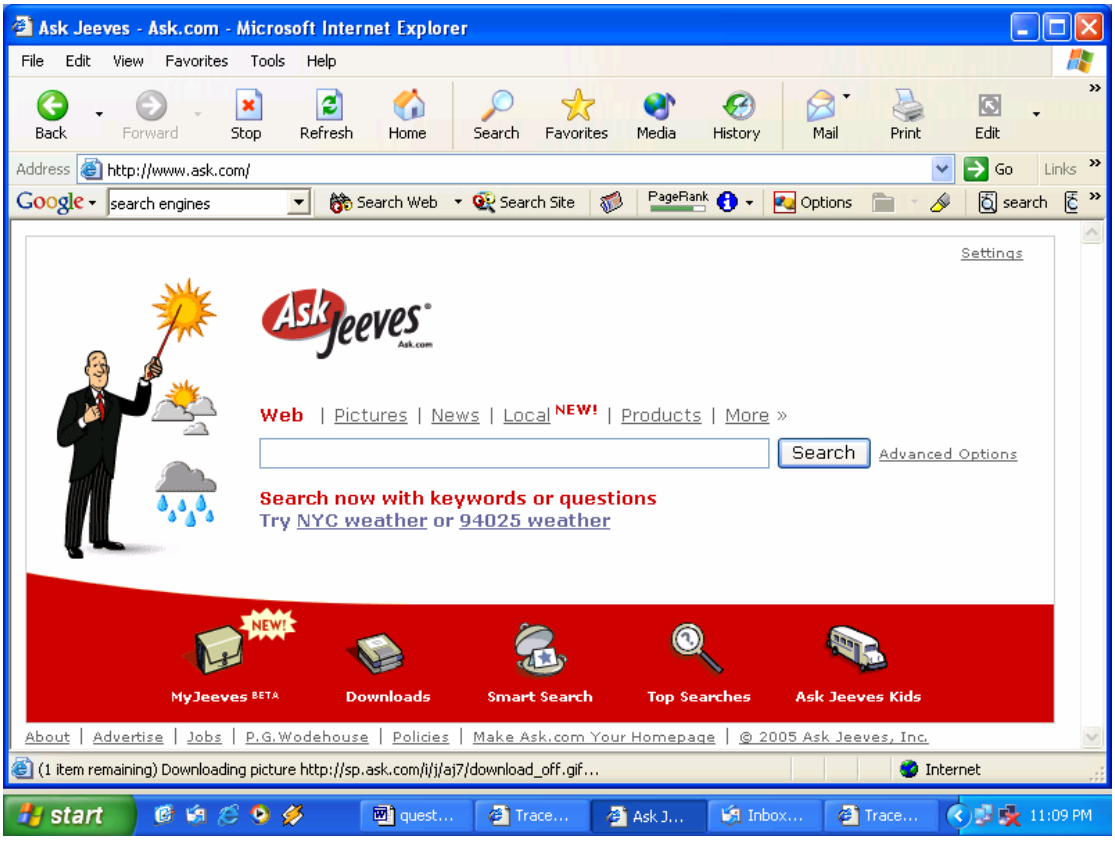





Excite - Web Search Home Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit
Address http://search.excite.com/
Go Links
Google search engines
Search Web Search Site PageRank Options search
excite
web search
Web Pages Images Audio NEW! Video NEW! News
Search
Exact Phrase
Advanced Web Search
Preferences
Tools & Tips
Web Directory
Directory Index
Autos
Careers
Computers
Entertainment
Fashion
Music
News
Real Estate
Shopping
Sports
Travel
EXCITE NEWS
The Excite Yellow Pages mean business!
From accountants to wedding planners, do your business online and save time every time -- with the Excite Yellow Pages. Just type, click and you're done.
Popular Searches
The Bachelorette
American Idol
Jen Schefft
50 Cent
Baby Names
Kelly Blue Book
Internet
start
quest...
Trace...
Googl...
Excit...
Lycos...
11:03 PM


AltaVista - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit
Address http://www.altavista.com
Go Links
Google
Search Web Search Site PageRank Options
AltaVista USA
altavista
Web Images MP3/Audio Video Directory News
FIND
Advanced Search
Settings
SEARCH: Worldwide USA RESULTS IN: All languages English, Spanish
Translate Toolbar Shopping Yellow Pages People Finder More ››
Business Services Submit a Site About AltaVista Privacy Policy Help
© 2005 Overture Services, Inc.
Make AltaVista My Home Page
Done
Internet
start
questionaire.doc ...
Yahoo! Search - ...
AltaVista - Micros...
10:59 PM


Lycos Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit
Address http://www.lycos.com/
Go Links
Google search engines Search Web Search Site PageRank Options search
LYCOS
Web People Discussion News Yellow Pages Products Multimedia
SEARCH THE WEB:
Go Get It!
Advanced Search | InSite Search Marketing
THE LYCOS 50 The Pop Culture Barometer
Movers & Shakers: Truth and Justice
What do Abraham Lincoln, Sandra Dee, and the Argentine actress Luciana Salazar all have in common?...more
NEW IN LYCOS SEARCH
YELLOW PAGES
Search the yellow pages. Find businesses
Lycos 50 Search Trends
2005 SI Swimsuit Models
1. Marisa Miller
2. Petra Nemcova
3. Yamila Diaz-Rahl
4. Carolyn Murphy
5. Ana Beatriz Barros
6. Jessica White
7. Anne V
NEW! Dating Search
Matches from multiple site
Millions of photo profiles from top c sites. More choice help you find your perfect match.
Monster
Find your next job on Mor
Internet
start
quest... Trace... Googl... Excit... Lycos...
11:03 PM


Ask Jeeves - Ask.com - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit
Address http://www.ask.com/
Go Links
Google search engines Search Web Search Site PageRank Options search
Settings
Ask Jeeves
Web | Pictures | News | Local NEW! | Products | More »
Search Advanced Options
Search now with keywords or questions
Try NYC weather or 94025 weather
NEW!
MyJeeves BETA
Downloads
Smart Search
Top Searches
Ask Jeeves Kids
About | Advertise | Jobs | P.G.Wodehouse | Policies | Make Ask.com Your Homepage | © 2005 Ask Jeeves, Inc.
(1 item remaining) Downloading picture http://sp.ask.com/i/j/aj7/download_off.gif...
Internet
start
quest... Trace... Ask J... Inbox... Trace...
11:09 PM