CHARACTERIZING WORLD WIDE WEB QUERIES

Ghaleb Abdulla, Binzhang Liu, Rani Saad, Edward A. Fox

Virginia Polytechnic Institute and State University

Blackburg, VA 24061-0106

{abdulla,liub,rsaad,fox}@vt.edu

ABSTRACT

Locating information on the WWW is a major activity for users, and Web Information Retrieval Systems (IRS) are becoming more important to support their endeavors. In this paper we characterize queries performed by Web users to such systems and give distributions for accesses to different Web IRS. We characterize clients' accesses, queries and user sessions. Our purpose is to reduce network and server bandwidth by identifying ways to optimize interactions with the Web. We characterize clients' sessions by a sequence of Browsing, Searching, and Next steps, and demonstrate that more search steps correlate with a reduction in the number of bytes transferred.

INTRODUCTION

The Web includes a huge distributed collection of books, articles, images and other kinds of media. Every day it grows either by adding new servers or by adding more information to the available servers. Information on the Web changes from day to day. Documents change because they need to be corrected, added to, or taken from. Locating information in this environment can be time consuming and frustrating. Users can spend hours while browsing, often looking for something without finding it.

To help users locate information on the Web, special search tools and cataloging systems were developed. We will call these tools Web Information Retrieval Systems (IRS). Unfortunately, users have not learned how best to use the newly developed tools. As a result many of their queries are not effective or efficient and they spend most of their time browsing the (returned) collection looking for relevant information. Borgman et al. [1] observed that "The end users who now dominate searching are using systems with exploratory interfaces, under less time pressure, and have less clear retrieval goals than do skilled search intermediaries."

The purpose of this paper is to characterize queries to Web IRS in an attempt to help with solving the Web scalability problem. This characterization is part of an ongoing effort to characterize different Web objects (e.g., users, networks, servers, and their subclasses) and how their methods and interactions can be optimized to help in reducing network and server workloads [2,3,4,5]. We offer suggestions for Web IRS designers and identify areas of training needed for Web users. We demonstrate that using Web IRS search can reduce network and server workloads.

The rest of the paper is organized as follows. The next section gives background information on current approaches to locate information using the Web. The third section covers our objectives; the fourth section describes the workloads studied; the fifth section characterizes workloads used; the sixth section characterizes queries; the seventh section characterizes client sessions. Finally, the last section presents conclusions and future work.

BACKGROUND

Currently users can look for information over the Web in one of three methods:

1- Browsing, users can follow links from page to page until they reach the information they are looking for. Most users get distracted during the browsing process and end up reading or looking at something completely irrelevant to their initial interest. However, browsing is still effective when users are looking for information within a specific collection that is limited in size. For example, consider locating a certain product or reading about a new product that is made by a well known company; in this case the URL for that company can be used to access their server and hence the desired page.

2- Using Web catalogues, like "Yahoo", users with a general topic can use the classification provided to narrow their subject until they find what they are looking for. For example, if we are interested in HTML and we want to read about it, then we can take the following path to get to the required information: from the main directory in the "Yahoo" home page to "Computers and Internet", "World Wide Web", and finally "HTML". Following these three steps will lead to several further choices about HTML including editing tools, standards, manuals, etc.

3- Using Web IRS, when looking for something more specific, users might turn to Web IRS like "Lycos" or "Infoseek" to help locate what they are looking for. For example, we may want to read a certain article and know the author of the article but not where this article was published. We can use the author's name and a key word from the article to locate it.

OBJECTIVES AND METHODOLOGY

The objectives of this paper are to:

1- Test the following four hypotheses:

Hypothesis 1: Users use Web IRS and catalogues to locate information.
If this is true then we will try to answer the following questions:

- What is the percentage of clients who used Web IRS in each workload?
- What is the distribution of queries/search engine (i.e, how many queries were done using "Lycos", "Infoseek", etc. for each workload)?

Hypothesis 2: A small percentage of the clients are responsible for most of the queries.

- This might not be true in a lab situation where students are instructed to find information about a certain topic (for example K-12 students during a lab session). If it is true, then what does this imply?
- What is the distribution of queries/client?

Hypothesis 3: Users do not use Web IRS effectively or efficiently.

Hypothesis 4 : The more searching we have, the less network bandwidth we use.

2- Characterize usage of Web IRS.
3- Pin down common mistakes and inefficient activities done by users.
4- Characterize clients' sessions in terms of "Browsing", "Searching", and "Next set of results" steps and study the relation with number of bytes transferred.
5- Characterize queries to a local Web IRS and highlight similarities and differences to activities on the Web in general.

WORKLOADS STUDIED

To test the previous hypotheses and answer the previous questions we will examine five log files that have been collected from different communities. We present a detailed description of the log files used in this analysis. The files were analyzed using scripts written in "Perl". Each script extracts the relevant data fields and parses them for specific information.

Table 1 describes the workloads used in this study, showing dates of log file collection and number of accesses. Each workload represents one or more log file from collected from the same place and concatenated together. An access is defined as a "GET" or "POST" HTTP command and is represented in one line in the access-log file of a proxy server. The workloads represent different groups of users with different ages and backgrounds so we can study differences between different groups and generalize our results whenever feasible.

The Korea proxy server log file contains accesses from all over the nation to servers inside and outside the country. It contains accesses from users in industry, education, and other communities. All these accesses have to pass through the proxy where they are logged. The workload from the Virginia Tech library contains accesses from clients inside the library to servers outside Virginia Tech. Accesses can be made from students who use machines in the Library hall or laboratory areas or from library employees. Accesses in the CS workload are mainly from graduate students in the department; in addition there are some accesses from the faculty. Most of the accesses logged are recorded on a machine that was used as a proxy server for other clients in the department and labs. The America On Line (AOL) workload contains accesses from AOL clients. The log file covers several hours from the original server access log file. For privacy reasons the log file contains names of accessed servers and URLs only. It does not contain date, time, or client information. However it is still of interest to us and it is a good source especially for testing hypothesis 1. The last workload was obtained from Auburn (AUB) high school near Blacksburg, VA. This workload was collected in a lab with high school students who were asked to locate information about a time varying event and learn how to represent it graphically.

All workloads were collected in 1996 except for the Korea workload which was collected in 1995. The table also shows the number of accesses in each workload. Based on results of this paper, we hope to get cooperation from other user communities or Web IRS operators to obtain additional log files of workloads.

Table 1 Workloads used in this study

Workload	Period	Number of (K) accesses
Korea	9/2-26/95	1682
Library (VT)	9/19-11/20/96	128
CS (VT)	10/9-11/10/96	92
AOL	11/96 (one day)	897
AUB	10/21-10/22/96	19

WORKLOAD CHARACTERIZATION

In this section we test Hypothesis 1. The following discussion is necessary to help clarify some of the tables and figures in subsequent sections. First, when we say "accesses to Web IRS" we include all accesses to Web IRS such as "Lycos", "Infoseek" and to cataloguing systems such as Yahoo. Second, we count the number of accesses to the search server including accesses that do not contain a search command and compare it to the total number of accesses in the workload. This way we compare total accesses with search servers to total accesses to all other servers.

CHARACTERIZING ACCESSES TO WEB IRS

Table 2 shows accesses to search servers and their percentages for each workload. We notice that in all workloads the percentage of accesses to Web IRS is no more than 20%. This is reasonable since the search activity is always followed by browsing activity to explore the returned URLs and how relevant they are to the search. However we also notice that in the high school case the percentage is very high compared to the others; this is because the students were instructed to use the Web to find information. Interestingly the library workload was higher than the rest which reflects a key service

of libraries, namely locating information. The Korea workload has the lowest percentage. One reason could be that the data was collected a year ago and Web users were not very well informed about the Web IRS. We expected that the CS workload ratio would be higher than other workloads in using Web IRS but CS students search the Web during only 6% of their accesses. This ratio is low compared to all other workloads except Korea. By examining the log file further we noticed that accesses from the CS students are mostly to servers inside the department and those accesses are mostly repeated accesses to pages that they are developing.

Table 2 Accesses to Web IRS and their percentages

Workload	Accesses to Web IRS	% of total accesses
Korea	67756	4
Library (VT)	15408	12
CS (VT)	5780	6
AOL	78026	9
AUB	3839	20

Table 3 lists the total number of clients in each workload and compares it with the number of clients who have accessed a Web IRS. We see that the ratio differs for different workloads. All clients in the high school have accessed Web IRS because they were supervised by teachers who knew about the available search tools and how to use them and who instructed the students regarding how to search.

Table 2 and Table 3 show that our first hypothesis is true. Web users do use Web IRS to locate information. However the degree of computer knowledge does not appear to predict the percentage of accesses that are queries, for our workloads.

Table 3 Total number of clients and clients which accessed a search engine

Workload	Total number of clients	Clients performing searches	%
Korea	2136	929	43
Library (VT)	348	203	58
CS (VT)	36	21	60
AUB	79	79	100

Table 4 shows the distribution of accesses to different Web IRS. One thing that is common between all workloads except AOL is that Yahoo is the top used search engine. It is not strange that Webcrawler is the top search engine used in the AOL workload since it is their official search engine. Why does Yahoo get so many accesses? It could be because it is one of the earliest Web directories and Web IRS, it has combined browsing and searching techniques, or because "Yahoo" is listed in the Netscape search page which is the page used by most of the users to access Web IRS. In a discussion with one of the users, she said that she uses "Yahoo" because the name is easier to remember and to type compared to, for example, "Altavista". A usability study might compare all these Web IRS and find out why one is used more than the others.

Table 4 Distribution of accesses to currently available Web IRS

Search Engine	Korea		Lib(VT)		CS(VT)		AUB		AOL	
	Acc.	%	Acc.	%	Acc.	%	Acc.	%	Acc.	%
Altavista	0	0	846	5	1073	19	20	1	3370	4
aol	193	0	1033	7	427	7	309	8	4235	5
elibrary	0	0	170	1	0	0	0	0	1420	2
Excite	4	0	2381	15	477	8	186	5	6842	9
Four11	31	0	185	1	35	1	46	1	1370	2
Infoseek	17361	26	1178	8	563	10	479	12	1588	2
Lycos	2811	4	2650	17	983	17	431	11	2463	3
Magillan	179	0	2015	13	277	5	406	11	71	0
nln	419	1	12	0	0	0	0	0	23	0
Opentext	312	0	29	0	0	0	0	0	48	0
Webcrawler	3653	5	155	1	97	2	473	12	46356	59
Yahoo	42792	63	4742	31	1834	32	1488	39	8028	10
100hot	0	0	0	0	14	0	0	0	60	0

Some of the Web IRS have 0 hits or 0 percentage hits; this does not mean that they are worse than the other tools. It could be that they are new and users have not discovered them yet. In addition some of these engines are specialized in finding people not information; as a result we expect them to be accessed less frequently than the other IRS. Nevertheless, for our purposes using data related to the varied workloads and the Web IRS shown in Table 4 should allow our results to be more generalisable than if a single workload or Web IRS was considered.

CHARACTERIZING CLIENTS' ACCESSES

In this section we study the distribution of accesses per client and the cumulative accesses for the clients in each workload. Figures 1-4 show the distribution of accesses for each client. In all workloads we notice that the distribution is "Weibull". We have a client which is very active followed by less active clients. The number of accesses decays rapidly and shows an asymptotic behavior. By fitting the data we found out that they follow a "Weibull" distribution with α=.44 for the Korea workload, α=.51 for the lib workload, α=.37 for the CS workload, and α=.48 for the AUB workload. This implies that a small percentage of the clients are responsible for most of the accesses.

Figure 1 Accesses/client distribution for the CS data

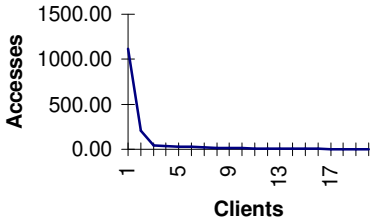


Figure 2 Accesses/client distribution for the AUB data

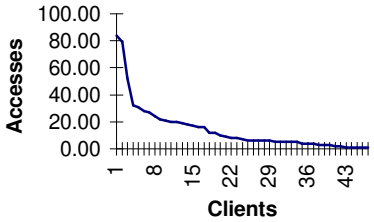


Figure 3 Accesses/client distribution for the Library data

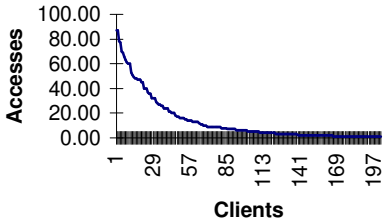
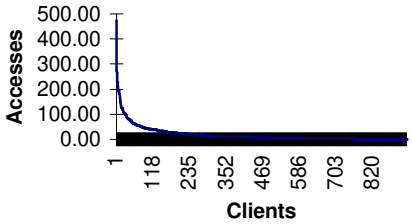


Figure 4 Accesses/client distribution for the Korea data



To test if this is true we plotted the percentage of cumulative number of accesses for the clients vs. the percentage of clients. Figures 5-8 show the results. In all graphs we see that 50% of the clients are responsible for more than 80% of the accesses. In some cases even 10%-20% of the clients are

responsible for more than 80% . What does this mean? It is hard to make conclusions about specific users' behavior since our data is about client machines. However we can make conclusions about general user behavior. We can say that a small percentage of clients do most of the search on the Web and the rest browse more than they search. Why is that? We speculate that users are not well trained to use Web IRS, or that they think of the Web as an extension to the TV where you browse channels until you find something interesting to watch. One area of training is to inform users about search tools and how to use them so more users can search.

Figure 5 Percentage of accesses vs. percentage of clients for the CS data

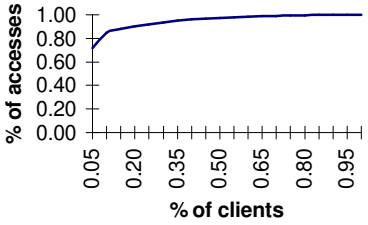


Figure 6 Percentage of accesses vs. percentage of clients for the AUB data

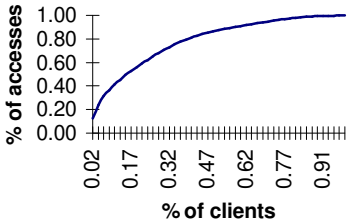


Figure 7 Percentage of accesses vs. percentage of clients for the Library data

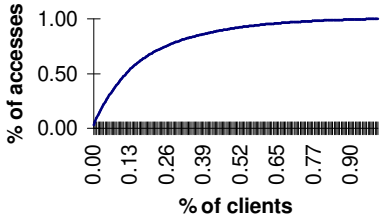
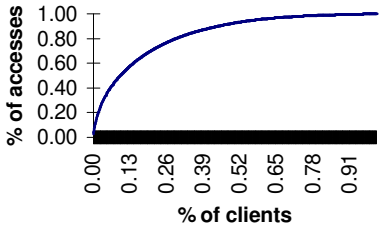


Figure 8 Percentage of accesses vs. percentage of clients for the Korea data



QUERIES

In this section we study queries submitted by users and also attempt to pin down the common mistakes and inefficiencies. We start by counting how many accesses to the Web IRS are queries and compare that with the total number of accesses to the Web IRS. Table 5 shows the number of accesses to Web IRS in each workload which contains a query. The rest of the accesses can be to load the home page of the Web IRS, to read the help about how to do the search, or to access inline images. The CS (VT) workload has the highest percentage. A higher percentage in this table means that the users in this workload accessed the Web IRS more to search and they did not browse as much. It can also mean that they favored accessing the IRS search interface over browsing its catalogue interface.

Table 5 Number and percentage of accesses that are searches

Workload	Number of queries	% of accesses
Korea	16845	22
Library (VT)	2515	16
CS (VT)	1503	26
AOL	5342	7
AUB	619	16

We next measure query complexity by counting the number of terms and number of operators used in queries. Figures 9-13 show the distribution of number of terms in the issued queries for each workload. From the figures we can see that users tend to use more than one term in their queries; in fact for all workloads except Korea (where perhaps search skills most need improvement) two terms/query is the most popular situation. When using more terms users may be trying to be specific to get more relevant results. In some queries the number of terms exceeded 5 terms. In a ranked system this is beneficial in spite of the large number of partial matches if good terms are chosen, due to statistical correlation, since users can define a threshold when they will stop looking at retrieved documents.

Figure 9 Number of terms/query, AUB

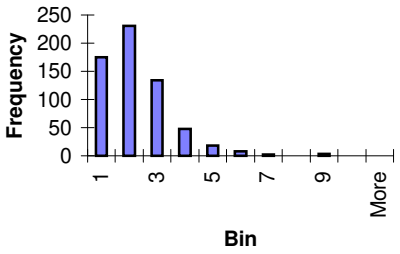


Figure 10 Number of terms/query, AOL

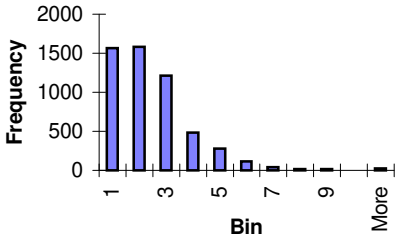


Figure 11 Number of terms/query, CS (VT)

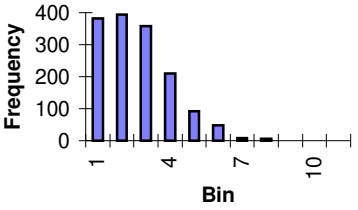


Figure12 Number of terms/query, Korea

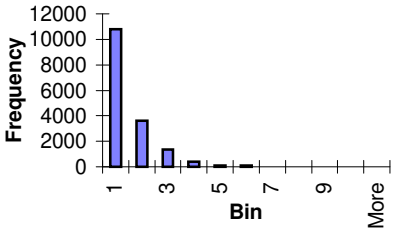


Figure 13 Number of terms/query, Library

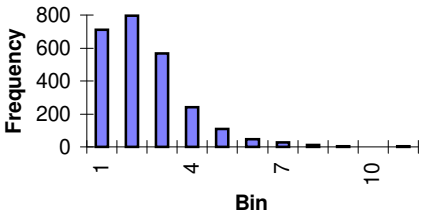


Table 6 shows the distribution of number of terms/query for each workload. In each case most of the queries had no operators, as would be expected for one-term queries, or queries to a "natural language" Web IRS. (Note: This is the case for the Web IRS that we have examined; all have a ranking feature, or allow Boolean queries if special syntax roles are followed.)

Table 6 Counts of number of operators/query

Workload	0 operators (%)	1 operator (%)	2 operators (%)
Korea	16672 (98.94)	172 (1.04)	1 (0.01)
Library (VT)	2408 (95.74)	105 (4.17)	2 (0.08)
CS (VT)	1492 (99.27)	11 (0.73)	0 (0.00)
AOL	5050 (65.32)	282 (33.49)	10 (1.19)
AUB	614 (99.19)	5 (0.81)	0 (0.00)

Table 7 shows the distribution of query types performed by users. "Advanced" stands for queries that use special features provided by the Web IRS. Again the table supports the conclusions from the previous graphs; users search most of the time using "Natural Language." Boolean or advanced queries are rarely used in any workload.

Table 7 Distribution of query types

Workload	NL (%)	Boolean (%)	Advanced (%)	Total
Korea	15172 (90.07)	317 (1.88)	1356 (8.05)	16845
Library (VT)	2409 (95.79)	71 (2.82)	35 (1.39)	2515
CS (VT)	1361 (90.55)	114 (7.59)	28 (1.86)	1503
AOL	5091 (95.30)	210 (3.93)	41 (0.77)	5342
AUB	597 (96.45)	11 (1.78)	11 (1.78)	619

Table 8 shows counts of some types of misunderstandings observed in queries submitted by Web users. Two major misunderstandings are use of Boolean operators and use of stop words: Do users think a natural language system can implement Boolean searching, or do they think stop words really help in a search (since users are generally reluctant to type, and few know how natural language systems work, we conjecture that their mental model involves computers understanding English.) Our numbers are of value since they provide an upper bound for such misunderstandings. In Korea's case the counts are low because the users used one term most of the time. In the other workloads the percentages are larger; this should be further investigated to find out the actual numbers of misunderstandings through user interviewing. Future work will count spelling errors, which inspection showed to be numerous.

Table 8 Counts of misunderstandings

Workload	Boolean (%)	Stop words (%)
Korea	74 (0.44)	289 (1.72)
Library (VT)	84 (3.34)	354 (14.08)
CS (VT)	101 (6.72)	30 (2.00)
AOL	196 (3.67)	543 (10.17)
AUB	72 (11.63)	5 (0.81)

HARVEST, THE SERVER MODEL

We collected data from the campus Web IRS server in the library at Virginia Tech accessible on the page www.vt.edu. This Harvest server is accessed by clients from inside and outside Virginia Tech. It represents a different model of data collection than all but our AOL workload since it is server based, not focused on the behavior of a group of clients as in the previous cases studied. The number of accesses to the server were 54118 out of which 1048 were invalid queries (1.94%); 12308 (22.74%) had 0 returned objects. The invalid queries were either due to spelling mistakes or misuse of operators. In addition Harvest users used stop words in 10.69% of the queries studied, indicating that they assumed it is a natural language system. However, Harvest is a Boolean system and does not rank returned objects, but handles a sequence of words by "adding" them.

Figure 14 and Table 9 show the distribution for number of terms and number of operators for users searching Harvest. The figures indicate that the distribution is essentially the same as the one noticed for other workloads. This may be true because the interface to Harvest is through a network browser so users may assume it is similar to other Web IRS.

.

Figure 14 Number of terms/query, Harvest

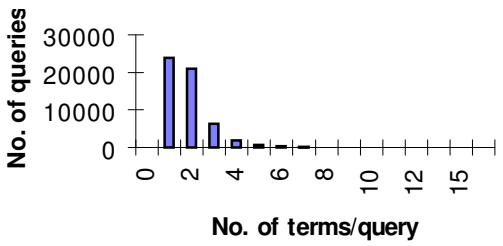
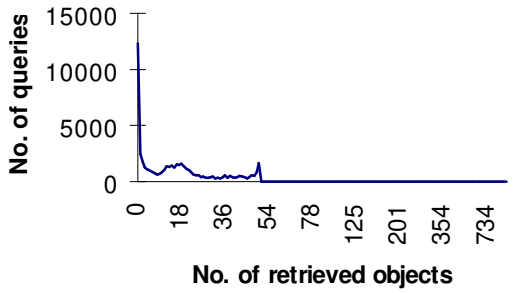


Table 9 Number of operators/query, Harvest

0 operators (%)	1 operators (%)	2 operators (%)	3 operators (%)	4 operators (%)	5 operators (%)	>5 operators (%)	Total
52700 (97.38)	1224 (2.26)	141 (0.26)	40 (0.07)	11 (0.02)	1 (0.00)	1 (0.00)	54118

Figure 15 shows the distribution for the number of retrieved objects from Harvest. We can see that most of the searches resulted in a small number of retrieved documents possibly because the collection searched is relatively small.

Figure 15 Number of queries vs. Number of retrieved objects



CHARACTERIZING SESSIONS

User and client sessions contain important information. The problem of extracting sessions is very hard and needs more research. In our study we will used heuristics to identify sessions and then studied the effects of number of searches on the total number of bytes transferred. We also examined the behavior of clients in terms of three steps: "browse" or "b", "search" or "s", and "next set of results" or "n". The first step in processing was to extract the interactions for each client from the log file that contains all accesses generated from that client. The next step was to identify session boundaries in each client's logfile. By examining the log files we estimated 300 seconds a safe number to indicate a session break. We scanned the client logs generating a sequence of interactions for each identified session. For example a certain client might have 10 sessions in its log file. Each session is represented with a line of "b", "s", and "n" codes for user actions.

In this section we test hypothesis four, examine some interesting session patterns and see if they are common between workloads or not. We define the following ratio (s+n)/b which is the ratio of search and next steps over the browsing steps. According to hypothesis four, if we have more search-related steps then the number of bytes transferred for the session should be less. We will use regression and correlation to determine the relation between the above ratio as well as counts of b, s, and n codes, and the number of bytes transferred.

The correlation analysis for four of the workloads (Korea, AUB, CS (VT), Lib (VT)) showed that there is a negative correlation between the ratio (s+n)/b when compared with the number of bytes transferred. This supports hypothesis four.

Let Tbytes be the total bytes transferred. Table 10 shows the regression results for the four workloads. The table contains the regression equation for each workload and under the equation we list the t statistics, in brackets, for each variable in the equation.

Table 10 Regression results for the four workloads

WorkLoad	Regression Results (t statistic) (t statistic)
Korea	Tbytes = 1050282 - 903312*(s+n)/b (19.33) (-5.49)
Lib(VT)	Tbytes = 836878 - 1077943*(s+n)/b (8.16) (-4.71)
CS(VT)	Tbytes = 1110089 - 1556183*(s+n)/b (14.35) (-5.38)
AUB	Tbytes = 1188716 - 4794527*(s+n)/b (4.47) (-1.95)

The regression results for all four workloads are very significant. The coefficients of the ratios (s+n)/b are all negative in four workloads. This negative coefficient suggests that a higher searching and browsing ratio means fewer total bytes transferred. In the Korea workload case, one percent increase in the ratio will cause a decrease by 9.03K; in the Library workload, it decreases by 10.78K; in the CS workload, it decreases by 15.56K; in the AUB workload, it decreases by 47.9K in a session. These results support hypothesis four.

The regression analysis for number of bytes transferred and the b, s, n counts are shown in Table 11. The coefficients of the variable b are significant in all workloads. The coefficients of the variable s are only significant in the library and AUB workloads. The coefficients of the variable n are only significant in Korea and library workloads. This suggests that b is the most important variable that determines the total bytes transferred.

Table 11 Regression results for bytes transferred and b, s, n

	Korea		Library (VT)		CS (VT)		AUB	
Variable	Coeff.	T ratio	Coeff.	T ratio	Coeff.	T ratio	Coeff.	T ratio
b	10929	38.46	3170	17.12	5483	19.35	9639	14.09
s	-2037	-0.13	128678	12.41	36788	1.83	-103431	-4.78
n	21519	2.05	46705	3.06	-47947	-1.68	-1084	-0.03

To analyze the access patterns, we replace a sequence of repeated b's with "B". Table 12 shows the top 10 popular access patterns workloads. The most popular access pattern in all sessions is the sequence with b steps only and it is represented by B in the table. This pattern occurs in over 85% of all sessions which suggests that users are browsing most of the time. It is interesting to note that the first three access patterns not only are common between all workloads but also that they have very close percentages. The second popular access pattern is BsB, which implies that the user has done one search after browsing and ended the session with browsing. Access pattern BnB is interesting because we had an n appearing without a search. By examining the log file we found out that this can be done if the user use the catalogue in some Web IRS to get to a certain category, then the user pressed next screen to get more hits. This pattern appears in all four workloads.

Table 12 The most frequently repeated string patterns in all workloads.

Access Pattern	Korea	Library (VT)	CS (VT)	AUB
B	85.25	93.66	89.78	85.25
BsB	5.63	3.86	4.02	5.63
BsBsB	2.00	0.62	1.72	2.00
BsBsBsB	1.15	0.20	0.70	1.15
BsBnB	0.51	0.05	0.14	0.51
BnB	0.46	0.93	0.27	0.46
BsBsBsBsB	0.35	0.07	0.41	0.35
BsBnBnB	0.30	0.01	0.00	0.00
BnBnB	0.15	0.08	0.15	0.15
BnBnBnB	0.15	0.03	0.00	0.00

CONCLUSIONS AND FUTURE WORK

In this study we found that users do use the Web IRS to locate and find information. We also found that small percentage of the clients are responsible for most of the accesses to the search engines. In particular less than 50% of the clients are responsible for more than 80% of the accesses. We found that the percentage of natural language queries is over 90% and users rarely use the additional features available through the Web IRS interface. By looking at the accesses patterns in the sessions we found that the most repeated pattern is browsing only. Finally, using regression and correlation analysis we found that the more searching, the less network and server bandwidth is used.

The Web is becoming an important source of information for millions of people; Web IRS should help users in locating information as fast as possible. This may be achieved by:

1- Developing interfaces that allow users to apply the most effective search methods faster.
2- Training users to use these systems and to perform efficient queries.
3- Training users to avoid pitfalls and misunderstandings and spelling mistakes.

Clearly more effective queries means less time to spend browsing and less bytes to be transferred. Web scalability is a major issue these days. Users started noticing delays to the extent that the name "World Wide Wait" is being used by some users to show their frustration with the Web's slow response. Our suggestions above should help.

A usability study should be done to identify why users rarely invoke advanced Web IRS features doing searches. More work should be done to characterize queries to Web IRS servers. Ultimately, through more log analysis, supplemented with user interviews and usability studies, better models of how users of the WWW find and work with information should enhance its scalability.

REFERENCES

[1] Christine L. Borgman, Sandra G. Hirsh, Hohn Hiller, "Rethinking Online Monitoring Methods for Information Retrieval Systems: From Search Product to Search Process," *Journal of the American Society for Information Science*, Vol. 47, No., July 1996.

[2] Marc Abrams, Stephen Williams, Ghaleb Abdulla, Shashin Patel, Randy Ribler, Edward A. Fox, "Multimedia Traffic Analysis Using Chitra95," Proceedings: *ACM Multimedia '95*, San Francisco CA, November 1995, pp 267-276.

[3] Marc Abrams, Charles R. Standridge, Ghaleb Abdulla, Stephen Williams, Edward A. Fox, "Caching Proxies: Limitations and Potentials," *Proceedings: 4th Inter. World-Wide Web Conference*, Boston, MA, Dec. 1995, pp 119-133.

[4] Marc Abrams, Stephen Williams, "Complementing Surveying and Demographics with Automated Network Monitoring," *World Wide Web Journal*, No. 3, Vol. 1., June 1996.

[5] Ghaleb Abdulla, "Characterizing and Modeling World Wide Web Objects", Ph.D. Dissertation in Progress.

[6] Christine L. Borgman, "Why Are Online Catalogs Still Hard to Use," *Journal of the American Society for Information Science*, Vol. 47, No., July 1996.

[7] Stephen P. Harter, "Online Information Retrieval Concepts, Principles, and Techniques," *Academic Press Inc.*, 1986, pp. 132-138.

[8] Averill M. Law, Stephen Vincent, "UniFit user's Guide", 1994.

CHARACTERIZING WORLD WIDE WEB QUERIES

Ghaleb Abdulla, Ph.D. student, Computer Science Department, Virginia Polytechnic Institute and State University, Virginia, 24061-0106, Phone: (540)231-3877, Fax: (540)231-6075, email: abdulla@vt.edu
Binzhang Liu, Graduate Student, Computer Science Department, Virginia Polytechnic Institute and State University, Virginia, 24061-0106
Rani Saad, Graduate Student, Civil Engineering Department, Virginia Polytechnic Institute and State University, Virginia, 24061
Edward A. Fox, Professor, Computer Science Department, Virginia Polytechnic Institute and State University, Virginia, 24061-0106

We would like this paper to be considered for the Best Student Paper Award.