Analysis of User Web Traffic with a Focus on Search Activities

Feng Qiu, Zhenyu Liu, Junghoo Cho
University of California
Los Angeles, CA 90095
{fqiu,vicliu,cho}@cs.ucla.edu

ABSTRACT

Although search engines are playing an increasingly important role in users' Web access, our understanding is still limited regarding the magnitude of search-engine influence. For example, how many times do people start browsing the Web from a search engine? How much percentage of Web traffic is incurred as a result of search? To what extent does a search engine like Google extend the scope of Websites that users can reach? To study these issues, in this paper we analyze a real Web access trace collected over a period of two and half months from the UCLA Computer Science Department. Our study indicates that search engines influence about 13.6% of the users' Web traffic directly and indirectly. In addition, our study provides realistic estimates for certain key parameters used for Web modelling.

1. INTRODUCTION

Since its arrival in the early 90's, the World-Wide Web has become an integral part of our daily life. According to recent studies, people access the Web for a variety of reasons and spend increasingly more time surfing the Web. For example, [1] shows that a typical Internet user spends more than 3 hours per week online and tends to spend progressively less time in front of the TV partly due to increased "surfing" time.

This research is motivated by our desire to understand how people access the information on the Web. Even though the Web has become one of the primary sources of information, our understanding is still limited regarding how the Web is currently used and how much it influences people. In particular, we are interested in the impact of search engines on people's browsing pattern of the Web. According to recent studies [2], search engines play an increasingly important role in users' Web access, and if users heavily rely on search engines in discovering and accessing Web pages, search engines may introduce significant bias to the users' perception of the Web [3].

The main goal of this paper is to quantitatively measure the potential influence of search engines and the general access pattern of users by analyzing a real Web access trace generated from the users' daily usage. For this purpose, we have collected all HTTP packets originating from the UCLA Computer Science Department from May 15th 2004 until July 31st 2004 and analyze it to answer the following questions:

- *Search-engine impact*: How much of a user's access to the Web is "influenced" by search engines? For example, how many times do people start browsing the Web by going to a search engine and issuing a query? How many times do people start from a "random" Web site? How much do search engines expand the "scope" of Websites that users visit?
- *General user behavior*: How many different sites do people visit when they surf the Web? How much time do people spend on a single page on average? How many links do people follow before they jump to a "random" page?

The answers to the above questions will provide valuable insights on how the Web is accessed by the users. Our study will also provide realistic estimates for some of the key parameters used for Web modeling. For example, the number of clicks before a random jump is one of the core parameters used for the *random-surfer model* and PageRank computation [4].

The rest of the paper is organized as follows. In Section 2 we describe the dataset used for our analysis. In Section 3 we report our findings on the influence of search engines on the users' Web access. In Section 4 we report our other findings on the general user behavior on the Web. Related work is reviewed in Section 5 and Section 6 concludes the paper.

2. DESCRIPTION OF DATASET

In this section we first describe how we collect our HTTP access trace and discuss the necessary cleaning procedures we apply to it to eliminate "noise."

2.1 HTTP access trace

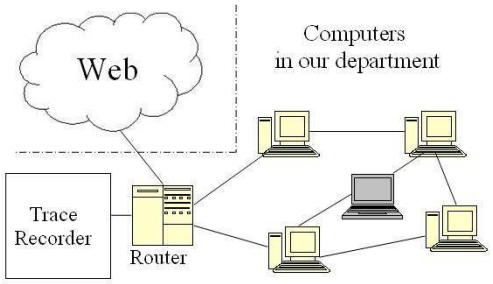


Figure 1: Network topology of UCLA CS Department

We have captured all HTTP Requests and Responses coming to/leaving from the UCLA Computer Science Department for the period of two and a half months. As we show in Figure 1, the CS department has roughly 750 machines connected through a 100Mbps LAN, which is then connected to the Internet through the department router. Since all packets that go to/come from outside machines pass this router, we can easily capture all HTTP packets by installing a packet recorder at the router. Given the large volume of traffic, we recorded only the relevant HTTP headers (e.g., Request-URL, Referer, User-Agent, etc.) in the packets, discarding the actual content.

Statistics	Value
Collection period	May 15th – July 31st, 2004
# of local IPs	749
# of remote IPs	66,372
# of requests	2,157,887
size of our trace (in bytes)	50GB

Table 1: Statistics on our dataset

To help the reader assess the scale of our HTTP trace, we report a few statistics of our dataset in Table 1. In brief, our dataset contains 2,157,887 HTTP Requests generated by 749 machines inside of our department while they access 66,372 outside servers over a period of two and a half months.[1]

2.2 Data cleaning

The goal of this paper is to understand the user behavior on the Web. Unfortunately, a significant fraction of our HTTP trace was due to various activities that are not directly relevant to user behavior (e.g., download traffic generated by Web crawlers). In this section, we describe three main filtering criteria that we use in order to remove the "non-user" traffic from our dataset.

- *Crawler traffic*: There are a few Web crawlers running in our department for a number of research projects. The traffic from these crawlers are clearly irrelevant to user behavior, but it constituted more than half of our collected data. We filter out this crawler traffic by discarding all packets coming from/going to a few machines where the crawlers run.
- *Non-textual traffic*: Users typically consider everything within a Web page (both text and images) as a *single* Web page; they do not consider images on a page as a completely separate unit from the surrounding HTML page. However, the browser issues multiple HTTP Requests to fetch embedded images, so if we simply count the number of HTTP Requests issued by browsers, there is a mismatch between what users see (*one* Web page) and what we count (say, *five* HTTP Requests). This mismatch is particularly problematic when a Web page contains many small icons or advertising banners.

  To avoid this mismatch, we decide to limit our analysis only to text documents (e.g., HTML, PDF, PS), because most non-textual objects are embedded in an HTML page and are perceived as a part of the page. That is, we keep only the HTTP Requests that retrieve textual documents. This filtering is done by checking the *Content-Type* field of the response for each request and keeping only those whose *Content-Type* value is "text/html," "text/pdf," etc.
- *Non-browser traffic*: A number of computer programs generate HTTP Requests that do not directly reflect the users' browsing behavior. For example, a *BitTorrent* client — a distributed content dissemination system [5] — generates frequent HTTP Requests to its neighbors to check their availability and to download files. Again, since our focus is on users' Web browsing behaviors, we eliminate the traffic from these clients by checking the *User-Agent* field of the requests and retaining only those requests from well-known browsers, such as "Mozilla."

Other than described above, we also eliminate certain obvious noises, like requests to URLs in wrong formats. Figure 2 shows the fraction of our original trace that is filtered out by each criterion described above. The crawler filtering is most significant; more than 60% of the traffic is discarded by this criterion. After the three filtering, we are left with 5.3% of the original trace, which is 2,157,887 HTTP Requests.

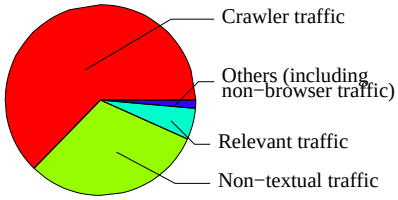


Figure 2: Fraction of discarded HTTP Requests

3. SEARCH ENGINE INFLUENCE

Based on the dataset described in the previous section, we now investigate how much search engines influence Web users. Search-engine influence can be seen from two different perspectives.

- *Help users visit more sites*: URLs of Web sites and/or pages are often hard to remember. *Bookmarks* or *Favorites* are used to maintain a user's favorite URLs, but they quickly become unmanageable as the list grows larger. Given this difficulty, users often use a search engine as an "extended bookmark"; they access pages by typing *keywords* (which are easier to remember) to a search engine instead of typing URLs. In this regard, search engines "expand" the set of pages that users can visit compared to the set of pages users have to remember or bookmark.

  How much do search engines expand the set of pages that a user visits? Is there overlap between the pages that users remember and visit directly and the ones that they visit through search engines?
- *Directing user traffic to particular sites*: Among billions of pages available on the Web, search engines direct users to a particular set of pages by picking and presenting a handful of pages in their search results given a query. Therefore, search engines "drive" a certain fraction of user traffic to the set of their selected sites. What fraction of user traffic is driven by search engines? How often do users randomly browse the Web and how often do they rely on search engines?

In order to answers the above questions, we first formalize "search-engine influence" by introducing the notion of a *referer tree*[2] in Section 3.1. We then present the statistics collected from our dataset in Section 3.2.

3.1 Influence, referer tree, and user

We assume that a user's visit to page p_2 is "influenced" by page p_1 if the user arrives at p_2 by following a link (or pressing a button) in p_1. This "link-click" information can be easily obtained from the *Referer* field in the HTTP Request headers. We illustrate the meaning of this field using a simple example.

Example 1 A user wants to visit the American Airlines homepage, but he does not remember its exact URL. To visit the page, the user first goes to the Google homepage (Figure 3(a)) by typing its URL `www.google.com` in the address bar of a browser. He then issues the query "American Airlines," for which Google returns the page

[1]The reported numbers are after we apply filtering steps described in the next section.

[2]In this paper, we use the misspelled word "referer" instead of the correct spelling "referrer" because of its usage in the standard HTTP protocol [6, 7].

in Figure 3(b). The user clicks on the first link and arrives at the American Airlines homepage (Figure 3(c)). From this homepage he further reaches other pages.

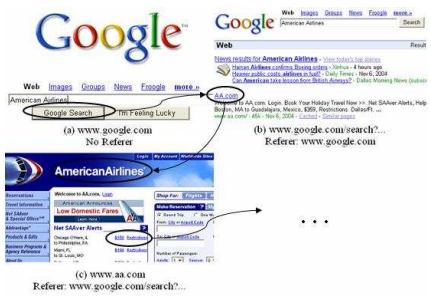


Figure 3: An example to illustrate the meaning of the *Referer* field

In this scenario, note that the user arrives at the first Google page (Figure 3(a)) directly without following a link. In this case, the *Referer* field of the corresponding HTTP Request is left empty, indicating that the user either directly typed the URL or used a bookmark entry. In contrast, the user arrives at the second and third pages (Figures 3(b) and (c)) by clicking a link or pressing a button. In these cases the *Referer* fields contain the URL of the immediately proceeding pages. For example, the *Referer* field of the second page request contains the URL of the first page, `www.google.com`. □

In summary, by looking at the existence and the value of the *Referer* field, we can tell whether and what links the user followed to arrive the page.

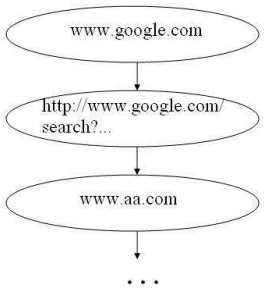


Figure 4: The referer tree for Example 1

Referer tree Using the *Referer* field information, we can construct a *referer tree*, where the nodes are the pages visited by a user and the edge from node p_1 to node p_2 means that the user followed a link from p_1 to p_2. In Figure 4 we show an example referer tree corresponding to the scenario described in Example 1. Note that the root of a referer tree represents a page visited directly by the user without following a link.

Given a referer tree, search-engine influence may be measured in one of the following ways:

- *Direct children only*: We consider that a search engine influences only the visits to the direct children of a search node (e.g., visit to www.aa.com node in Figure 4). This interpretation is reasonable in the sense that the search engine cannot control the links that its direct children present to the user.
- *All descendants*: We consider that all descendants of a search node are under search-engine influence. This interpretation is also reasonable because if the search engine did not provide the link to its direct children, the user wouldn't have arrived at any of their descendants.

In the next section, we estimate search engine influence under both interpretations.

Users In order to analyze users' Web browsing behaviors, we need to associate every HTTP Request with an individual user. In general, automatic user identification of an HTTP Request is a very complex task [8]. Fortunately, the usage pattern of our department machines allows us to use a simple heuristic for this task with reasonably high accuracy: we assume that *each IP corresponds to one user*, because all faculty members and most students have their own workstations that they primarily use for accessing the Internet.

The only concern is that some IP addresses might correspond to server machines, not workstations. On one hand, some of the servers are time shared; multiple users may simultaneously access the Web from a server, so the requests from one server represent the *aggregate behaviors* of multiple users, not the behavior of a single user. On the other hand, many servers are primarily used for computational tasks and practically no user uses them to access the Web. Therefore, if we count the requests from these servers in computing user statistics, the results may be biased.

To avoid these problems, we rely on the fact that more than 90% of user workstations run Windows or Mac operating systems, and consider only the requests from those machines when we try to measure the behavior of individual users.

3.2 Results of search-engine influence

We now report our results on search-engine influence. We notice that more than 95% of the search activities from our department goes to three major search engines: Google, Yahoo! Search and MSN Search. For this reason, we primarily focus on the influence of these three search engines in the rest of this section.

Search-engine-directed traffic In Figure 5, we show the fraction of traffic to search engine home pages (e.g, the first level nodes in Figure 4), to search engine result pages (e.g., the second level nodes in Figure 4), to their direct children (e.g., the third level nodes in Figure 4) and their descendants. Roughly, 1.0% of the user traffic goes to search engine home pages, and 5.7% are search requests. 2.1% of the user traffic goes to the direct children of search requests, with additional 4.8% to their descendants. (For this set of reported statistics, we have excluded the set of search-engine-home-page loading requests that do not lead to any further traffic, since such requests are most likely results of setting search engines as the default loading page of a Web browser.) Overall, 13.6% of user traffic is under the direct and indirect influence of search engines. Interestingly, these results imply that many of our users issue queries to search engines but do not click on links in the result pages.

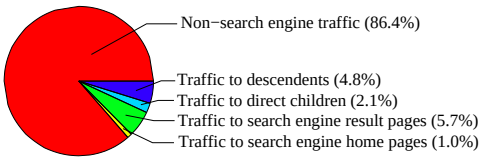


Figure 5: Search-engine traffic size

We can also assess the search-engine influence by measuring how many times people start surfing the Web from search engines. Given that the root node of a referer tree is where a user starts his surfing, we can measure this number by counting the number of search-engine-rooted referer trees. Our dataset contains a total of 380,453 referer trees, out of which 25,758 are rooted at search engines. Thus, we estimate that in about 6.8% of the time our users start surfing the Web from search engines.

Helping users visit more sites We now discuss how much search engines expand the set of sites that a particular user visits. This "site expansion" by search engines can be viewed in two ways: (1) search engine increase the *number of "starting points"* from which users can browse further (by providing new links in its search results.) (2) search engine increase the *total number of sites* that the user eventually visits. We may estimate these two effects of search engines as follows:

- *Seed-set expansion*: We refer to the set of Web sites from which a user starts his Web surfing as the *seed set* of the user. Given this definition, the *regular seed set* of a user corresponds to the root nodes of her referer trees (except when the root node is a search engine). The *search-engine seed set* corresponds to the direct children of search engine nodes.That is, the set of sites that search engines refer to, from which the user starts browsing. We can measure the seed set expansion by search engines simply by comparing these two seed sets.
- *Visit-set expansion*: We refer to the set of sites that a user eventually visits as the *visit set* of the user. The *search-engine visit set* is the set of all descendants of search-engine nodes. The *regular visit set* is all the nodes in the referer trees except the search-engine descendants. Again, by comparing these two sets, we can measure the visit set expansion by search engines.

In Figure 6, we first plot the seed set expansion effect by search engines. In the figure, the horizontal axis corresponds to time and the vertical axis shows the sizes of the regular seed set, the search-engine seed set and the overlap between them after the given time interval. For example, after six weeks, an average user has 188.2 sites in his regular seed set and 56.7 sites in the search-engine seed set, with an overlap of 15.6 sites. We observe that the relative ratio of this overlap roughly remains constant over the period of 10 weeks, which is about 8% of the regular seed set. We also observe that search engines consistently expands the size of the seed set by 22% over this period of time.

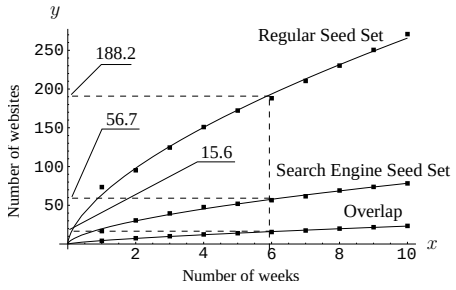


Figure 6: Search-Engine Seed Set Expansion

In Figure 7, we show a similar graph for the visit-set expansion. The meaning of the two axes of this graph is similar to the previous one. After six weeks, a user visits a total of 246.0 sites without using search engines and 72.4 sites starting from search engines, with an overlap of 23.0 sites. Similarly to the previous seed-set results, the relative size of the overlap roughly remains constant at a level of 9% of the regular visit set. Overall, search engines help an average user visit 20% more sites and the sites that users visit through search engines seem quite distinct from the sites that users visit from random surfing.

4. USER ACCESS STATISTICS

In this section we try to measure users' general behaviors in surfing the Web. In particular, in Section 4.1 we investigate how an average user follows hyperlinks during Web browsing. In Section 4.2 we investigate how much time people spend per page and how long they stay online in "one sitting."

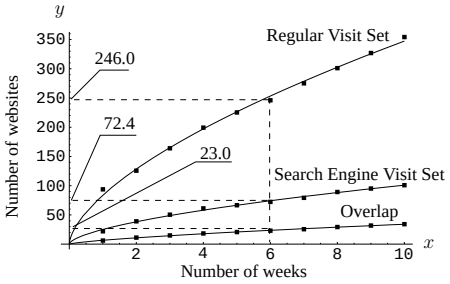


Figure 7: Search-Engine Visit Set Expansion

4.1 Referrer tree statistics

Hyperlinks are considered a core structural component on the Web. It is generally believed that hyperlinks play a significant role in guiding people to particular Web sites. Since users' actions of following links are fully captured by referer trees, we now analyze the characteristics of the referer trees in our trace to understand the our users' clicking behavior.

In particular, we are interested in the *size*, *depth* and *branching factor* of the referer trees. The size of a referer tree measures how many pages a user visits by following links before she jumps to a new page. The depth shows how deeply a user follows hyperlinks before she stops exploring further. The branching factor indicates how many links on a page a user typically clicks on.

In Figure 8, we show the distributions of these three properties. In the graphs, the horizontal axis corresponds to the size, depth and branching factor of refer trees, respectively. The vertical axis shows the number of referer trees with the given characteristics.[3] All graphs in this section are plotted in a log-log scale.

From the graphs, we first see that all distributions closely fit power-law curves; the graphs are straight lines in the log-log scale. Also from Figure 8(a), we observe that 173,762 out of 380,453 referer trees have a single node. That is, 45% of the time, users jump to a completely new page after visiting just one page. Finally, given the mean of each distribution[4] we estimate that a typical Web user visits 5 pages by following hyperlinks, clicking on 3 links per page, but going down no more than 3 links deep.

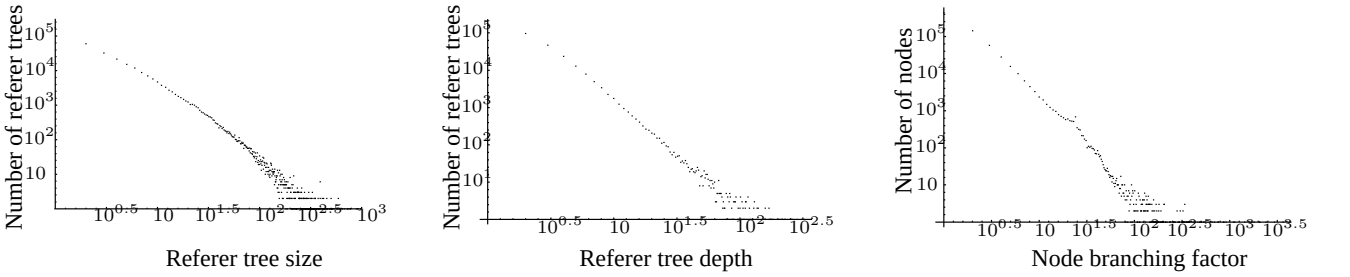
4.2 Session statistics

We now report statistics on the following characteristics of user behavior: (1) How many pages and sites do people visit once they start surfing the Web? (2) How much time do they stay online in one sitting? (3) How many times do they jump to new pages while they surf the Web? In order to answer these questions, we first introduce the notion of a *session*.

Definition of session Informally, a session refers to the set of pages that a user accesses in one "sitting." A traditional definition for the session is based on time-out. That is, after a user starts visiting Web pages, if there is a certain period of inactivity, say 10 minutes, then the current session expires and a new session starts. The main weakness of this definition is the difficulty in choosing a good time-out value. On one hand, if the time-out is set too short, the pages that a user browses in one sitting may be broken into multiple sessions, especially if the user reads a long online article. On

[3] More precisely, because branching factors are characteristics of individual *nodes* not of *trees*, Figure 8(c) shows the number of nodes with the given branching factor.

[4] In computing the average branching factor, we exclude the leaf nodes in the tree for which users did not click any links.



(a) Size distribution (avg:5.67) (b) Depth distribution (avg: 2.52) (c) Branching factor distribution (avg: 2.95)

Figure 8: Referer tree characteristics

the other hand, if the time-out is set too long, the pages that the user accesses in multiple sittings may be combined into one session. To remedy this shortcoming, we decide to extend the traditional definition using the referer-tree information.

The basic idea for our extended definition is that even if a user accesses a page after a certain period of inactivity, if the user clicks on a link on a previously accessed page to access a new page, it strongly hints that the user was actually reading the previous page. Based on this intuition, we put the accesses to page p_1 and p_2 into one session

- if they are accessed within a short time interval τ or
- if p_2 is accessed by following a link in p_1.

For example, consider Figure 9 that shows a sequence of pages accessed by a user. The relative spacing between the pages represent the time interval elapsed between the accesses. The curved arrows at the top represent that the user followed a link in the first page to the second. In this example, (p_1, p_2, p_3), (p_4, p_5), and (p_6, p_7, p_8) are put into the same sessions because they are accessed within time τ. In addition, p_3 and p_4 are put into the same session because p_4 is accessed by following a link in p_3. Overall, pages p_1 through p_5 are put into one session and pages p_6 through p_8 are put into another session.

We believe that it is safe to use a small threshold τ value under our extended definition, because as long as the users follow a link to reach from one page to another, these two pages are put into one session, even if the access interval is longer than τ. For this reason, we use a relative small value for τ, 5 minutes, for our analysis.

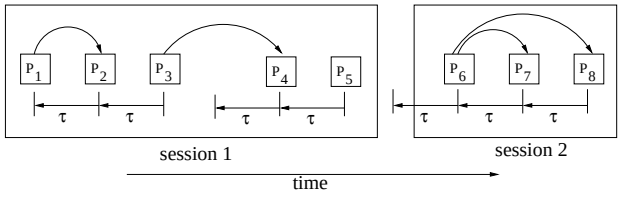


Figure 9: An example of our method of identifying sessions

Number of Web sites and pages per session Based on the session definition given above, we first report how many Web pages and sites a user visits in one session. In Figures 10 and 11 we present the distributions for Web pages and Web sites per session, respectively. The horizontal axis corresponds to the number of pages (or sites) per session, and the vertical axis shows the number of sessions that have the given number of pages (or sites). The average numbers are 21.79 for Web pages and 5.08 for Web sites, which means that a typical user visits about 22 pages in 5 Web sites in one sitting. The graph for Web pages closely fits a power-law curve, while the graph for Web sites does not exhibit a close fit.

Session length and average time per page Another interesting statistics is how much time a user spends online once she starts

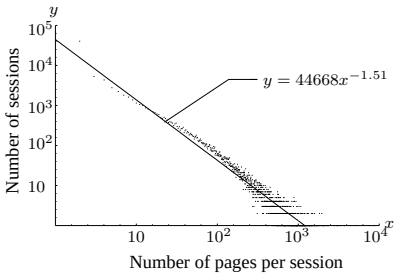


Figure 10: Number of pages per session (Avg:21.79)

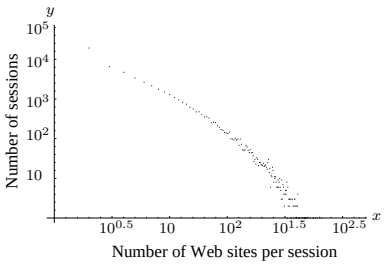


Figure 11: Number of Web sites per session (Avg:5.08)

surfing the Web and how much time she spends reading each page. One issue in measuring these numbers is how to account for the time spent on the last page of a session. Because there is no subsequent page access, we do not know when the user stops reading the page. As a rough approximation, we assume that the time spent on the last page is equal to the average time spent on a Web page. Based this assumption, we present the session length distribution in Figure 12 and the average time per page distribution in Figure 13.

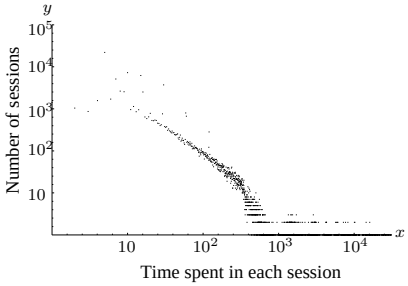


Figure 12: Session length in time (units in minutes)

The graphs have a large number of outliers, but the general trends fit well to power-law curves. On average, a typical Web user spends about 2 hours per session and 5 minutes per page.

Number of referer trees per session Finally, we report within a session, how many times a user stops following links and jumps to

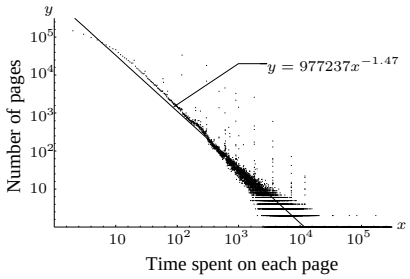


Figure 13: Time spent on each page (units in seconds)

a random page (by typing in a new URL or selecting a bookmark). Note that whenever the user jumps to a new page, a new referer tree is initiated. Thus we can learn how many times a user jumps to a random page in a session by counting how many referer trees the session contains. We present the number-of-referer-trees-per-session distribution in Figure 14. Again, the curve fits well to a power-law curve. The mean of the distribution is 3.83, meaning that people make about 3 random jumps per session on average.

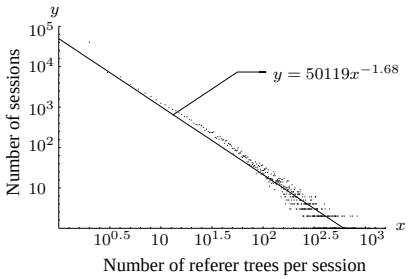


Figure 14: Number of referer trees per session (Avg:3.83)

5. RELATED WORK

Researchers have studied cognitive and behavioral aspects of user's Web search activities in the past [9, 10, 11, 12, 13, 14]. In these studies the main focus is how various subjective factors such as users' information need, knowledge, expertise and past experiences affect users' search behavior. Researchers also attempt to build cognitive or behavioral models (e.g. state transition graphs) to explain such behavior. In contrast, our study mainly focuses on *quantifying* the influence of Web search in people's daily Web access.

Our work is also related to earlier studies on how users surf the Web by following static links [15, 16, 17]. Compared to these studies we emphasize more on users' search behavior.

There has also been extensive research in general characteristic of Web queries [18, 19, 20]. A rather comprehensive review of such studies can be found in [21]. While these works mainly focus on reporting the statistics of Web queries by inspecting search engine logs, in this paper we are more concerned about the impact of search activities by studying Web search in a larger context of user's overall Web access.

6. CONCLUSION

In this paper, we tried to provide a quantitative picture on how users access the information on the Web using a 2.5-month Web trace data collected form the UCLA Computer Science Department.

We summarize some of our main findings as follows:

- We find that about 13.6% of all Web traffic is under the direct or indirect influence of search engines. In addition, search engines help users reach 20% more sites by presenting them in search results, that may be otherwise unreachable by the users.
- A typical Web user follows 5 links before she jumps to a new page, spending 5 minutes per page. In one sitting, she visits 22 pages residing on 5 Web sites.

One limitation of our study is that our observation was made on a potentially-biased user population. Therefore, some of the characteristics that we observed may not be generalizable to the entire Web user population. While we believe our quantification methods to derive such characteristics extend easily, it will be a interesting future work to see how some of our observations may change for a larger user population.

7. ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. 0347993. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

[1] SIQSS: Internet and society study, detailed report. http://www.stanford.edu/group/siqss/Press_Release/internetStudy.html, 2000.

[2] Brian Morrissey. Search guiding more Web activity. http://www.clickz.com/news/article.php/2108921, 2003.

[3] J. Cho and S. Roy. Impact of Web search engines on page popularity. In *Proc. of WWW '04*, 2004.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proc. of WWW '98*, 1998.

[5] BitTorrent. http://bittorrent.com/.

[6] World Wide Web Consortium. Hypertext Transfer Protocol - HTTP/1.1. www.w3.org/Protocols/ rfc2616/rfc2616.html, 1999.

[7] World Wide Web Consortium. HTML 4.01 specification. www.w3.org/TR/html4/, 1999.

[8] P. Baldi, P. Frasconi, and P. Smyth. *Modeling the Internet and the Web*, chapter Modeling and Understanding Human Behavior on the Web. WILEY, 2003.

[9] C. Hoelscher. How Internet experts search for information on the Web. In *Proc. of WebNet '98*, 1998.

[10] C. Holscher and G. Strube. Web search behavior of Internet experts and newbies. In *Proc. of WWW '99*, 1999.

[11] C.W. Choo and B. Detlor. Information seeking on the Web: An integrated model of browsing and searching. *First Monday*, 5(2), 2000.

[12] R. Navarro-Prieto, M. Scaife, and Y. Rogers. Cognitive strategies in web searching. In *Proc. of the 5th Conf. on Human Factors & the Web*, 1999.

[13] A. Broder. A taxonomy of Web search. *SIGIR Forum*, 36(2), 2002.

[14] D.E. Rose and D. Levinson. Understanding user goals in Web search. In *Proc. of WWW '04*, 2004.

[15] L.D. Catledge and J. Pitkow. Characterizing browsing strategies in the World-Wide Web. *Comp. Networks ISDN Syst.*, 27:1065–1073.

[16] L. Tauscher and S. Greenberg. Revisitation patterns in World Wide Web navigation. 1997.

[17] A. Cockburn and B. McKenzie. What do Web users do? an empirical analysis of Web use. *Int. J. Human Computer Studies*, 54:903 – 922.

[18] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large Web search engine query log. *SIGIR Forum*, 33(1):6 – 12, 1999.

[19] B.J. Jansen, A. Spink, and T Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing and Management*, 36(2):207 – 227, 2000.

[20] A. Spink, B.J. Jansen, D. Wolfram, and T. Saracevic. From E-Sex to E-Commerce: Web search changes. *IEEE Computer*, 35(3):107 – 109, 2002.

[21] B.J. Jansen and U. Pooch. A review of Web searching studies and a framework for future research. *JASIST*, 52(3):235 – 246, 2001.