A Characterization of Online Browsing Behavior

Ravi Kumar
Yahoo! Research
701 First Avenue
Sunnyvale, CA 94089.
ravikumar@yahoo-inc.com

Andrew Tomkins*
Google, Inc.
1600 Amphitheater Parkway
Mountain View, CA 94043.
atomkins@gmail.com

ABSTRACT

In this paper, we undertake a large-scale study of online user behavior based on search and toolbar logs. We propose a new *CCS taxonomy* of pageviews consisting of Content (news, portals, games, verticals, multimedia), Communication (email, social networking, forums, blogs, chat), and Search (Web search, item search, multimedia search). We show that roughly half of all pageviews online are content, one-third are communications, and the remaining one-sixth are search. We then give further breakdowns to characterize the pageviews within each high-level category.

We then study the extent to which pages of certain types are revisited by the same user over time, and the mechanisms by which users move from page to page, within and across hosts, and within and across page types. We consider robust schemes for assigning responsibility for a pageview to ancestors along the chain of referrals. We show that mail, news, and social networking pageviews are insular in nature, appearing primarily in homogeneous sessions of one type. Search pageviews, on the other hand, appear on the path to a disproportionate number of pageviews, but cannot be viewed as the principal mechanism by which those pageviews were reached.

Finally, we study the burstiness of pageviews associated with a URL, and show that by and large, online browsing behavior is not significantly affected by "breaking" material with non-uniform visit frequency.

Categories and Subject Descriptors. H.3.m [**Information Storage and Retrieval**]: Miscellaneous

General Terms. Experimentation, Measurement

Keywords. Browsing, Toolbar analysis, Pageviews

1. INTRODUCTION

Since the inception of the World-Wide Web some fifteen years ago, the rate of appearance of new capabilities, data types, and services has remained high, allowing users to meet more of their communications, entertainment, and task completion needs online. Every day, new websites appear, exploring new approaches and business models in competition with existing online and offline alternatives. Hundreds of millions of users engage daily with social networking sites that offer capabilities that were unknown just a few years ago. Search of webpages, videos, images, commercial listings, personal ads and so forth continues to grow in volume and sophistication. Users engage with the offerings afforded by this crucible, and their behavior evolves as rapidly as the business landscape underlying it. This evolution, however, is difficult to observe, partly because of its velocity and partly because user behavioral data is not generally available for study. Thus, we lack an accurate picture of how users engage with the Web.

In this paper, we perform a broad study of online user behavior based on data collected through the Yahoo! toolbar. We study a large sample of over fifty million user pageviews collected over a seven-day period in March of 2009 from users who have installed the Yahoo! toolbar and agreed to collection of their data for purposes including this type of analysis. (We will discuss the forms of bias that may be introduced by employing toolbar data for our study.)

Main findings. Our primary result, which will inform most of the analyses we perform, is a newly-proposed *CCS taxonomy* of online pageviews whose three top-level branches are *content* (news, portals, games, verticals, multimedia), *communication* (email, social networking, forums, blogs, chat), and *search* (web search, item search, multimedia search). We will define these in detail, and provide an additional one or two levels of depth to the taxonomy. Our summary finding is that roughly half of all pageviews online are content, one-third are communication, and the remaining one-sixth are search.

Our development of the CCS taxonomy is based on an editorial labeling of a random sample of pages. With the taxonomy in place, we then develop a series of automated recognizers for pageviews that lie in certain classes of the taxonomy. We employ these larger-scale sets of labeled data to study how users navigate within and among different types of pages, how search interacts with other types of navigation, and how users revisit pages within each taxonomy category. We show that mail, news, and social networking pageviews are insular in nature, appearing primarily in homogeneous sessions of one type. Search pageviews, on the other hand, appear on the path to a disproportionate number of pageviews, but cannot be viewed as the principal mechanism by which those pageviews were reached.

Finally, we study the burstiness of pageviews associated with a URL. We consider a "smoothed" variant of the observed data, in which each pageview appears throughout the time of our measurements according to an estimated Poisson

*This work was performed while the author was at Yahoo! Inc.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

likelihood that maintains overall frequency while removing all burstiness from the data for each URL. We show that the inter-arrival distribution of this smoothed model appears almost unchanged from that of the original data, allowing us to conclude that "breaking" material is not a significant contributor to the structure of online page visit frequency.

Organization. The remainder of the paper is organized as follows. Section 2 describes related work, and Section 3 characterizes the data sets we study. Section 4 gives some statistical characterization of users, sessions, time online, inter-arrival distributions, and popular destinations. Section 5.1 presents our taxonomy of pageviews, along with a series of analyses based on this breakdown. Section 6 defines a notion of a search session and studies user behavior in such sessions. Finally, Section 7 contains concluding remarks.

2. RELATED WORK

The related work falls into three main categories: modeling and studying general web browsing activity, modeling search and search-related activities, and toolbar-based analysis of user browsing activity.

Web browsing activity has been extensively analyzed and modeled in the literature. One of the earliest is the work of Catlege and Pitkow [9], who used both client- and server-side data in order to study web browsing behavior. Their study predates the existence of many of the current search engines and social applications; see also [11]. Montgomery and Faloutsos [26] identified various browsing trends and user browsing patterns. Bucklin and Sismeiro [7] considered the "stickiness" of a website, which is a factor responsible for repeat visits. They used server-side logs in order to analyze patterns of repeat user visits. Park and Fader [29] examined cross-site user visit behavior and proposed a multivariate timing model to use information from one site to explain the behavior at another. Johnson et al. [18] studied search and browsing behavior across e-commerce sites that are competing in nature. Website revisitation is an often revisited research theme: some early papers include [28, 16, 32]. Recently, Adar et al. [1] examined the relationship between the content change on webpages and people's revisitation patterns.

The work closest to ours in terms of defining a browsing taxonomy is that of Morrison et al. [27]. They proposed a simple taxonomy based on user's response to what web activities impacted their actions — the taxonomy has three dimensions, namely, purpose of search, method of search, and the type of content sought. Their analysis was small-scale, human-based, and predates modern web-based applications. Recall that a well-known and highly-cited taxonomy already exists for web search — the work of Broder [5].

Downey et al. [13] introduced a language based on state-machine representation for describing searching and browsing behavior on the client side. This provides a unified framework for analyzing general models of user behavior, including many server-side models that were proposed earlier [12, 22, 19, 20, 30], and constructing machine-learned models in order to predict the next action of the user. Mei et al. [25] proposed an analogous and general analysis framework for modeling search-related activities; see also [15, 10], who introduced a Bayesian click model for relevance. There have been several papers on modeling user interaction with search engines. Lau and Horvitz [22] introduced a Bayesian network to predict topic transitions in query logs by considering the context of the query along with inter-query time period. Radlinski and Joachims [30] identified sequences of queries on the same topic using features based on shared words in the queries. Spink et al. [31] addressed the problem of topic switching and multi-tasking in query sessions.

With the increased availability of user browsing data via the toolbar, toolbar log analysis is blossoming into an active research area. Büchner et al. [6] looked at the problem of pattern discovery from user navigation; visualization aspects of such navigation patterns were considered by Cadez et al. [8]. Mayr [24] developed a quantitative measure called the web entry factor to aggregate common usage frequencies for webpages, where an entry means a website visit with an identifiable entry pattern (navigation type) from a logfile perspective. Using the browsing data, Liu et al. [23] proposed BrowseRank, a serious enhancement to PageRank, which takes into account the time spent on webpages. They employ a continuous-time Markov chain in order to model their stochastic process. Downey et al. [14], Bilenko and White [3], and Bilenko et al. [4] tracked the browsing behavior after the user departs the search engine and begins to follow an information thread through the Web. In particular, they explored the connection between the information goal of the users and their search and navigation patterns.

3. DATA

In early 2008, Yahoo! began asking users at toolbar installation time if they would give permission for Yahoo! to log their pageviews. For users who give permission, all pageviews are logged.

We consider a random sample of users drawn from Yahoo! toolbar logs over a one week period from March 18, 2009 to March 24, 2009. The number of pageviews in the sample for each day are shown in Figure 1. We also gathered data for March 25, 2009, which we employ using the previous week of data as historical context. Match 18 and March 25, 2009 fall on a Wednesday.

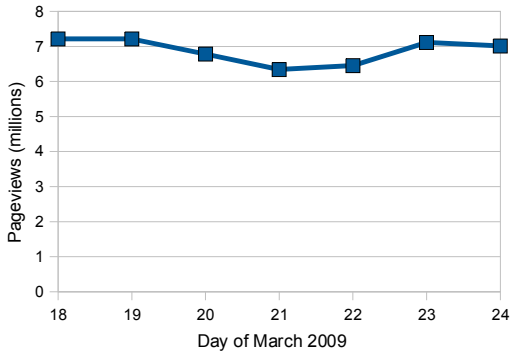

Figure 1: Number of pageviews in toolbar sample.

Toolbar biases. Before we proceed to analyze the data, we mention the caveat of the following biases introduced by sampling from Yahoo! toolbar data.

- Yahoo! toolbar penetration is significantly higher in the United States than elsewhere in the world.

- Users of Yahoo! toolbar typically install the toolbar in order to access some feature of the toolbar, such as mail notifications, or search. In addition, users with the sophistication to install a toolbar will behave differently from other users. These are examples of a variety of selection biases that may exist in studying toolbar users.
- Even if the user installed the toolbar in order to access a service she is already familiar with, the presence of the toolbar may lead to engagement with related services, or deeper engagement with the target service.
- Yahoo! toolbar began asking users for permission to log data in January of 2008. Users who have not upgraded their toolbar since that date are not included in the sample.
- Statistics regarding usage of Yahoo! properties are not representative of the general population.

4. BASIC USAGE

In this section, we present a quantitative analysis of the mechanics of online behavior. We begin in Section 4.1 by characterizing the distribution of pageviews and time online, at the level of both users and sessions. In the same section, we also characterize the inter-arrival time between pageviews within a session.

4.1 User and session characteristics

We begin by looking at the properties of overall usage of users. We have already discussed the biases resulting from use of toolbar data. In this section, we must be aware of one additional bias: users whose online behavior is sufficiently infrequent that a week may pass without any activity may not be represented in our sample at all.

We begin with a characterization of the daily time spent online and the number of pageviews. In order to perform this, we must estimate the time spent online during a particular pageview. Clearly, the time to load the page is an underestimate, but the time until next pageview is an overestimate. We address this issue by adopting a standard notion of a *session*. A sequence of pageviews for a particular user is broken into subsequences called sessions between any two consecutive pageviews whose timestamps are more than thirty minutes apart. Thirty minutes is a commonly used threshold for breaking sessions; see [31] for some discussion. We assume that time during a session is spent online, while time between sessions is not online.

Based on this decision, Table 1 characterizes the distributions of the number of pageviews, and of the total time spent online. The top part of the table presents information aggregate by user on a daily basis. The bottom part of the table shows information about sessions.

The second row of each section shows the median values. For users, the median number of pageviews per day is 59, and the median time online is about an hour. The other information in the table characterizes the tails of the distribution. 1% of users spend more than 9 hours per day online, and view over 927 pageviews averaged over our sample. Likewise, 10% of users spend only three minutes per day and view fewer than 5 pages.

Individual sessions tend to be significantly smaller. The median length is 17 pageviews taking 16 minutes of time.

Per user			
Total pageviews per week		Total time online per week	
Pageviews	% users	Total time	% users
≤ 3	11	≤ 30 sec	10
≤ 113	50	≤ 1.6 hrs	50
≤ 1097	90	≤ 15 hrs	90
≤ 3614	99	≤ 41 hrs	99
Per session			
Total pageviews		Total time online	
Pageviews	% users	Total time	% users
≤ 2	13	≤ 7 sec	10
≤ 17	50	≤ 17 min	50
≤ 113	90	≤ 1.5 hrs	90
≤ 385	99	≤ 4.0 hrs	99

Table 1: Total pageviews and total time online for users and sessions.

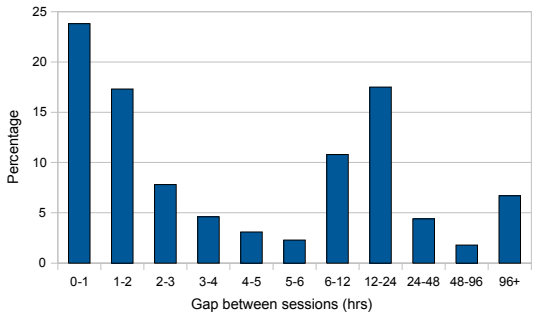

Figure 2: Gaps between sessions of the same user.

The tails of the distribution are likewise reduced. This suggests that users must engage in multiple sessions per day. The following table characterizes the distribution of the number of sessions per user across the eight days of our dataset:

Sessions per user	
# sessions	% users
≤ 1	32
≤ 4	53
≤ 25	90
≤ 48	99

Finally, Figure 2 gives the distribution over the gaps between sessions of the same user; we have provided slightly more detail on this distribution, as it is key to understanding how users interact online. 70% of sessions start within twelve hours of the previous session, and only 13% of sessions occur after a gap of a day or more.

To close this discussion of mechanical properties of search, Table 2 gives some information about the gaps that occur between pageviews within a session. The median pageview has a gap of only 12 seconds, and fewer than 10% of pageviews have a gap that stretches to more than 90 seconds. On the lower end of the distribution, around one in six pageviews have less than a second inter-arrival time. Some key factors contributing to short inter-arrival times are rapid use of the back button, but more importantly, multi-window and

Inter-arrival time	% pageviews
≤ 1 sec	16
≤ 11 sec	50
≤ 1.5 min	90
≤ 12.8 min	99

Table 2: Inter-arrival time between successive pageviews within a session.

multi-tab browsing. We do not have information in our logs to distinguish between these causes.

4.2 Popular websites

Table 3 gives the top sites under two different measures. The first measure is the overall number of pageviews from that site. The second is the total number of sessions that contain a URL from that site. The perspectives given by the two measures are quite different.

Before we delve into the data analysis, we remark that a careful canonicalization of the host name has to be done in order not to miss some popular websites. First, to eliminate noise, we ignore any URL matching the following patterns: `login`, `casalmedia`, `googleads`, `rd*.yahoo.com`, and `adserving.com`. Next, we hand-craft explicit rules for Yahoo!, Google, MSN, `pogo.com`, and `globo.com` to canonicalize specific servers into a generic one; for example, our rules will transform `mail301.yahoo.com` to the canonical `mail.yahoo.com`.

First, we observe that social networking sites such as Facebook, Myspace, Orkut, Friendster, Hi5, and Tagged, have achieved a pre-eminent position among the "head sites" online. This is particularly surprising, considering how recently these sites arrived as major players on the online scene.

Second, we observe that, while total pageviews are largely dominated by these sites, the session measure contains significant numbers of search and portal sites like MSN, Yahoo!, and Google. Users tend to have long-running sessions in which they engage deeply with social networking sites, performing large numbers of pageviews. However, while these sessions are longer, there are fewer of them than the shorter sessions searching, checking mail, or catching up on news.

Rank	Host pageviews	Host sessions
1	`facebook.com`	`google.com`
2	`myspace.com`	`youtube.com`
3	`google.com`	`facebook.com`
4	`orkut.com`	`mail.live.com`
5	`youtube.com`	`hotmail.com`
6	`friendster.com`	`myspace.com`
7	`hi5.com`	`images.google.com`
8	`craigslist.org`	`microsoft.com`
9	`tagged.com`	`en.wikipedia.org`
10	`ebay.com`	`mail.google.com`
11	`images.google.com`	`msn.com`

Table 3: Top sites by number of pageviews (left) and number of sessions referencing a URL from that host (right). References to Yahoo! hosts have been removed as toolbar data can be misleading.

5. PAGE-LEVEL ANALYSIS

In this section, we develop a taxonomy of basic pageview types, and characterize user behavior across this taxonomy. We present results of an editorial evaluation of pageviews with respect to this taxonomy. Next, we describe a set of automated recognizers that let us perform larger-scale analyses on pages from certain key nodes of the taxonomy. In the context of these automated recognizers, we then revisit users and sessions. We consider the extent to which users revisit the same page, and the interaction between page type and revisit frequency. Finally, we study how users move between pages, and examine link following within and across different types of page.

5.1 Taxonomy of pageviews

We adopt the following basic taxonomy of page types.

CONTENT. Any pageviews focused on a particular topic or area:

- NEWS: Pageviews providing news. This includes videos and photos meant to provide news.
- MULTIMEDIA: Pageviews delivering multimedia (audio, video, image) to the user. Image and videos on news sites are counted as news.
- PORTAL: Entry points allowing users access to a wide range of vertical content, such as Yahoo! or MSN.
- HEAD LISTINGS: Popular websites that contain listings, such as Amazon, Ebay, and Craigslist.
- GAMES: Pageviews providing access to online games of any form.
- ADULT: Pageviews providing adult content.
- OTHER VERTICAL: Vertical sites not covered in the above. See below for a more detailed breakdown of this category.

COMMUNICATION. Pages primarily focused on interacting with other users, often through sending or receiving some form of message:

- MAIL: Pageviews at email providers.
- SOCIAL: Pageviews at social network sites such as LinkedIn, Facebook, Twitter, Myspace, Hi5, Bebo, etc. Also includes smaller social networks, such as those dedicated to a particular geographic area or organization. To distinguish from groups, forums, or blogs, we view social networking sites as organizing content primarily based on user, as opposed to a forum that is organized principally by threads, or a blog that is organized primarily by message date.
- BLOG: Pageviews on blogs of any form.
- FORUM: Pageviews on forums or groups or online chats of any form.

SEARCH. Search result pages, of various forms:

- MAIN SEARCH: Pageviews on web search sites such as Google, Yahoo!, MSN Live, AOL search, and Ask. Restricted to pageviews that are part of web search, rather than other types of search.

MULTIMEDIA SEARCH: Multimedia search, for example provided by Youtube, Hulu, and the multimedia search URLs of the engines described above.

ITEM SEARCH: Search through a database of listings, as provided by the search boxes of Amazon, Ebay, or Craigslist.

Figure 4 gives overall pageviews for the categories shown above. 52% of total pageviews come from our broadly construed CONTENT category. Of these, GAMES represent 6% of overall pageviews, and MULTIMEDIA around 5%, mostly in the form of Youtube videos. PORTAL pages, at 5%, are typically entry points and browser homepages representing a gateway to mail, search, or verticals. HEAD LISTINGS covering Ebay, Craigslist, and Amazon provide about 3%, as does NEWS. These together account for about a quarter of total web pageviews. Another quarter come from a wide range of vertical sites, broken out in Table 5.

COMMUNICATION pageviews represent 35.5% of the total. Surprisingly, SOCIAL pageviews represent almost 25% of total pageviews. These are dominated in our sample by Myspace and Facebook, but the other large social networking sites such as Orkut, Friendster, Hi5, and Tagged also provide significant numbers of pageviews. As expected, MAIL is also a significant fraction of total pageviews, with around 13% of total. MULTIMEDIA sites, BLOG, and FORUM are less significant in overall pageviews.

SEARCH result pages represent 9% of total pageviews, which is much larger than we anticipated. Below, we will extend this evaluation to include the pageviews resulting directly or indirectly from some type of search. SEARCH is dominated by standard web search (MAIN SEARCH), although MULTIMEDIA SEARCH is significant.

As a rough caricature, one may imagine that half of pageviews are CONTENT, one-third are COMMUNICATION, and one-sixth are SEARCH.

The notion of MAIN SEARCH, ITEM SEARCH, and MULTIMEDIA SEARCH appeared in an earlier work [21], but that work considered only activities within search; this work covers all online pageviews.

Main category	Sub-category	Fraction
CONTENT	GAMES	6.2
	MULTIMEDIA	5.4
	PORTAL	5.4
	HEAD LISTINGS	3.4
	NEWS	3.4
	OTHER VERTICAL	28.1
	Total	**52.0**
COMMUNICATION	SOCIAL	24.3
	MAIL	9.4
	FORUM	1.4
	BLOG	0.4
	Total	**35.5**
SEARCH	MAIN SEARCH	6.2
	MULTIMEDIA SEARCH	1.4
	ITEM SEARCH	1.4
	Total	**9.0**
UNKNOWN	Total	**3.4**

Table 4: Pageviews broken by page type.

Vertical	Fraction
UNKNOWN	6.2
RETAIL	3.4
TRAVEL	1.8
FINANCE	1.4
EDUCATION	1.2
PERSONALS	1.0
JOBS	1.0
SERVICES	1.0
B2B	1.0
SOCIAL	0.8
ENTERTAINMENT	0.8
MOBILE	0.8
REFERENCE	0.8
SPORTS	0.8
REAL ESTATE	0.6
MOVIES	0.6
AUTO	0.6
TV	0.6
LOCAL	0.6
RADIO	0.4
FOOD	0.4
HEALTH	0.4
GOVERNMENT	0.4
CLASSIFIEDS	0.4
FASHION	0.2
WEATHER	0.2
NONPROFIT	0.2
MUSIC	0.2

Table 5: Breakout of OTHER VERTICAL pageviews from Table 4.

5.2 Automated analysis of page types

Many types of analyses we would like to perform with respect to the taxonomy can only be performed if we are able to classify significantly more pages than our manual study would allow. Thus, we develop a set of automated recognizers that approximate classification of pages into the taxonomy.

5.2.1 Simple recognizers

Here, we describe the nodes for which we have built recognizers, and we outline the (simple) recognizer algorithms themselves. (The recent work [2] on purely URL-based classification is applicable here.)

MULTIMEDIA: Custom recognizers for Youtube, Hulu, Blinkx, Flickr, Photobucket, and Smugmug.

PORTAL: Custom recognizers for entry points for Yahoo!, MSN, and Google.

HEAD LISTINGS: Custom recognizers for Ebay, Amazon, and Craigslist.

SOCIAL: Custom recognizers for the top 23 social networking sites.

MAIL: We recognize hostnames containing "mail" as a component of the name; this captures almost all the large email hosting organizations with good precision — we did not observe any false positives in our experiments.

MAIN SEARCH: Custom recognizers for the five largest US search engines: Yahoo!, Google, MSN, Ask, and AOL.

MULTIMEDIA SEARCH: Custom recognizers for multimedia search result pages from Yahoo!, Google, MSN, Ask, AOL, Youtube, Hulu, Flickr, and Picasa.

ITEM SEARCH: Custom recognizers for the URLs corresponding to search results from these listings providers: Amazon, Ebay, Craigslist, IMDB, Singlesnet, Careerbuilder, and Leboncoin.

OTHER SEARCH: For URLs not matching any of the recognizers above, we add a set of general rules matching any URL that has a search-like parameter embedded in the URL, such as `&q=madonna` or `&search=madonna`. We refer to these pageviews as OTHER SEARCH.

UNKNOWN: All other pageviews.

First, we examine how the performance of the recognizers compares to the results of the hand labeling. The results are shown in Table 6. Overall, 51% of pageviews are labeled by our automatic recognizers. We tuned the recognizers for high precision, and attained recall simply by including the most important sites or by identifying key patterns; we discuss below the bias that might be introduced towards head sites. The table shows that the recognizers are highly effective at identifying pages in the given classes. In the COMMUNICATION class, SOCIAL is within 2%, and MAIL is within 10%. In CONTENT, MULTIMEDIA catches just over 50% of the total multimedia pages, largely because multimedia content has a heavy tail and our recognizer catches only the head multimedia sites. PORTAL catches over 75% of pageviews, and HEAD LISTINGS catches 65%. For SEARCH, MAIN SEARCH catches 80% of pageviews; MULTIMEDIA SEARCH is almost identical, and ITEM SEARCH catches only 1/3 of pageviews, as this category also contains a heavy tail.

Main category	Sub-category	Fraction
CONTENT	MULTIMEDIA	2.8 (5.4)
	PORTAL	4.1 (5.4)
	HEAD LISTINGS	2.2 (3.4)
	Total	**52.0**
COMMUNICATION	SOCIAL	24.6 (24.3)
	MAIL	9.4 (8.6)
	Total	**35.5**
SEARCH	MAIN SEARCH	5.1 (6.2)
	MULTIMEDIA SEARCH	1.5 (1.4)
	ITEM SEARCH	0.5 (1.4)
	OTHER SEARCH	1.7 (0)
	Total	**9.0**

Table 6: Fraction of pageviews labeled by automatic recognizers for each category. Fractions from editorial study shown in parentheses for comparison. Table shows only categories for which we have an automatic recognizer.

Our recognizers for SOCIAL, MAIL, MAIN SEARCH, and MULTIMEDIA SEARCH all capture in excess of 80% of the pageviews for their category, and hence any bias towards head sites is bounded. MULTIMEDIA and PORTAL capture in excess of 50% of in-category pageviews, so while there might be some bias, and we believe that the interactions on head sites for these categories would be considered by most researchers to be exemplary of the category. The results for ITEM SEARCH are minimal, and should be viewed as representative of the particular sites we recognize, rather than the category at large.

5.3 Session reuse

To begin with, we employ these recognizers to determine which fraction of sessions and users contain pageviews of a given type. As we have seen earlier, and see again here, this number can be quite different from the overall pageview distribution given above. Table 7 shows the results.

Category	% sessions	% users
UNKNOWN	84.7	92.7
PORTAL	57.7	72.9
MAIL	42.7	62.3
MAIN SEARCH	33.9	62.6
SOCIAL	22.5	36.3
OTHER SEARCH	13.3	42.1
MULTIMEDIA	8.6	25.0
MULTIMEDIA SEARCH	6.7	20.8
HEAD LISTINGS	3.7	11.4
ITEM SEARCH	1.7	6.2

Table 7: Fraction of sessions and users viewing a certain page category.

Having characterized how often users view URLs of a particular type, we can now ask whether the particular URLs being viewed are new and different ones, or are simply the same places again and again. We employ the following methodology to study this issue. For the seven day period from March 18, 2009 through March 24, 2009, we record the latest occurrence of each URL. We then scan each URL visited on March 25, 2009 and determine whether it was viewed by the same user during the previous week. We say that the pageview is a *revisit* if the same user visited the same URL at some other point during the previous week. We say that the pageview is a *long revisit* if the same user visited the same URL during the previous week, but did not visit the URL in the previous 24 hours. Studying long revisits in addition to revisits allows us to control for the effect of revisits during a protracted session.

Table 8 shows the results. As a first observation, a significant number of pageviews (31%) have been seen by the same user during the previous week, and 2/3 of those have been seen in the last day. Notably the PORTAL category has 82% of pageviews as revisits. While the magnitude of the number is surprising, we would expect portals to score high on this measurement as the dominant portal strategy today is to create "entry points" such as the Yahoo! or MSN homepage that are carefully programmed to appeal to repeat visitors.

All forms of search underindex in both revisits and long revisits. We expect this, as search URLs contain embedded queries, and hence are more likely to be unique. Nonetheless, while such URLs are much more likely to be unique than in other categories, over a longer period of time, work of Teevan et al. [33] shows that a significant fraction of search queries involve the task of "refinding" a URL that a user has already searched for.

5.4 Referrals

We have covered our taxonomy of page types, and whether users visit distinct or already-seen pages. We now look at

Pageview category	% revisits	% long revisits
All	31	12
Head listings	28	13
Item search	45	13
Multimedia	23	8
Multimedia search	25	2
Mail	16	6
Main search	39	8
Other search	21	5
Portal	82	69
Social	28	11
Unknown	34	13

Table 8: Fraction of pageviews with revisits and long revisits, broken by pageview category.

how users navigate to these URLs, by considering the source of the link that took a user to a particular page. 35% of pageviews have no such link — the user arrived at these pages from bookmarks, direct typing into the URL bar, links from other applications, and so forth. We can therefore break all pageviews into two classes: first, there are "starting points," which do not contain a referrer field, and second there are "referrals," which do contain a referrer field of the referrals, we break out a set of referral types based on the page types defined above. Each pageview can be assigned to a referral class as follows.

- If there is no referrer for the page, assign class Starting Points;
- else if there is a referrer whose type is a subclass of search, assign the class of the referrer;
- else if the host of the referrer is the same as the host of the current page, assign the class Same-site Links;
- else if the referrer's type is not unknown, assign the class of the referrer;
- else assign the class Off-site Links.

Table 9 shows the relative proportions of each class of referral types. Same-site Links and Starting Points between them capture almost 80% of all referrals, and the remainder are almost always Search. Mail, with 10% of total pageviews, is responsible for almost 1% of referrals. And Social, with 25% of pageviews, is responsible for only 0.1% of (off-site) referrals. We observe that this finding raises some questions about recent excitement around mining social network sites as a source of outgoing links. It is possible that the 0.1% of referrals from these sites are of extremely high quality, interest, and timeliness, but they do not represent significant volume.

5.4.1 Referral inter-arrivals

Earlier, we discussed the distribution of the inter-arrival times between successive pageviews. In the context of referrals, we can also describe the distribution of the times between loading a page, and following a link from that page. We might expect these inter-arrival times to be shorter, as they represent the subset of page visits in which the user is more heavily engaged. On the other hand, we might expect

Referral type	% pageviews
Same-site Links	44.5
Starting Points	34.4
Off-site Links	11.5
Main search	5.3
Other search	1.5
Multimedia search	1.4
Mail	0.9
Item search	0.6
Social	0.1

Table 9: Breakout of pageviews by referral type.

them to be longer as the referrer to a page might not be the previous page in the user's session, due to artifacts like tabbed browsing. Table 10 shows the results: inter-arrival gaps from referrer to referred pages tend to be significantly longer than gaps between sequential pages in a session. The median is 50% larger, and the 90th percentile is larger by a factor of 3.

Inter-arrival time	% pageviews
$\leq$ 0 sec	14
$\leq$ 18 sec	51
$\leq$ 5.2 min	90
$\leq$ 3.1 hrs	99

Table 10: Inter-arrival time between referring and referred URLs.

5.4.2 Referral forests

At this point, we have presented some results about single-step referral chains. In an earlier work [21], we defined a referral forest for a user session, and studied how search queries appeared in these forests. We recap the definition, and study now how referral forests interact with the CCS taxonomy.

The *referral forest* of a session has a tree for each starting point, and a edge from each node to its referrer. As there can be many starting points within the same session, the resulting object may be a forest rather than a single tree.

The first question we ask is the following. Given that a user has arrived at a certain page, which types of pages appeared along the path taken to reach the given page? Table 11 gives these results. The values for search URLs are reprinted from [21]; the remaining values are new. Notice that the fraction of pageviews reached from Social or News pages is largely unchanged from their fraction in the global population. This implies that these types of pageviews are largely insular, in the sense that users do not typically navigate from these types of pages to other types of pages. The fact that the proportions are no smaller than the underlying distribution suggests that there are non-trivial sessions that contain almost exclusively pageviews of this type.

This analysis answers the question "Does a page of given type appear in the path taken by the user to arrive at a given URL?" Again following [21], we may define two mechanisms for dividing "credit" for a pageview among the ancestors of that pageview. The *root measure* assigns all credit for a pageview to the root of its referral tree. The *path measure* assigns credit to each ancestor such that the parent gets

Page type	Fraction
Unknown	51
Social	26
Main search	16
Mail	12
Other search	4.4
Multimedia	4.1
Multimedia search	3.3
Head listings	2.3
Item search	0.9

Table 11: Fraction of pageviews reachable from pageviews of a given type.

some credit, the grandparent gets half as much, and so on up the tree to the root.

Table 12 shows the results. Notice first that the root and path measures are almost identical, giving us some faith that the mechanisms are robust. Second, the results should be read in comparison to Table 4. Our initial expectation was that Search and to a lesser extent Mail would be a key source by which users discovered new content. In fact, the fraction of pages reached through Search, Mail, and Social sites is similar to their fraction in the overall population. This suggests that, while there is likely to be a search pageview on the path to interesting content (per Table 11), such pageviews are a stepping stone along the way, rather than the sole attributable provider of access to key content.

Page type	Root measure	Path measure
Unknown	38	45
Social	25	26
Portal	13	3.9
Mail	11	10
Main search	7.7	5.9
Multimedia	3.2	3.2
Head listings	1.8	1.9
Other search	1.1	1.9
Multimedia search	0.5	1.7
Item search	0.1	0.4

Table 12: Results for different measures of assigning credit transitively to referrers.

5.5 Burstiness

In this section, we study in more detail the behavior of multiple users either gradually or suddenly beginning to examine a particular URL. We start with some simple modeling. Figure 3 shows the cumulative distribution of the inter-arrival times between any visit to a URL and the previous visit (across all users simultaneously) over the course of a single day. The figure also shows the result of fitting this cumulative distribution with a logarithmic function of inter-arrival time; as one can see, the fit is so good that the curves almost entirely obscure one another. The actual form of the fit is

$$P(x) = 0.025 \cdot \log(x) + 0.366.$$

One may interpret this fit to mean pageviews fall into roughly equal-mass buckets of doubling inter-arrival times: 1–2 sec, 2–4 sec, etc. A natural explanation would be that

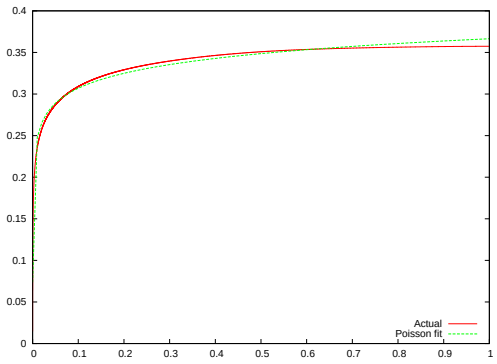

Figure 3: Actual inter-arrival time distribution and fit to a logarithmic function.

pages fall into equal-mass buckets of popularity: 0.01–0.02 views/sec, 0.02–0.04 views/sec, and so forth.

This leads to two hypotheses of bursty behavior on the Web. In the first model, each URL has a "popularity" that determines how frequently it is visited. This popularity changes only slowly, so in the course of a month, we model it as constant. In this model, a URL that receives a certain number of visits over the course of a month receives them spread out in a memoryless fashion, rather than clumped up during a single busy period.

Another hypothesis, however, would suggest that this misses one of the important properties of the Web: on any given day, there is always content that is "breaking" on that day, whether it be an official news story, a video of a dancing baby, or a quirky website referenced on slashdot or Yahoo!. Characterizing all pages as receiving traffic uniformly over time will be sufficiently inaccurate that the first model will perform poorly.

Figure 4 evaluates these two hypotheses by performing the following thought-experiment. First, in order to cover a broader range of inter-arrival times for analysis, we expand our dataset to cover four weeks of activity. For this month of time, we gather for each URL the total number of visits to the URL. We then assume that a URL with 10 visits over the course of 28 days contributes 10 pageviews with inter-arrival time drawn from a Poisson distribution with rate $28/10 = 2.8$ days. Likewise for all other URLs. This produces a set of pageviews equal to the number of pageviews in the original datasets, each with an inter-arrival time drawn from a Poisson model with an appropriately-chosen rate. One may view this strawman model as capturing the exact structure of the original data, except that all burstiness has been removed from the data, and URL arrivals have been smeared uniformly over the course of the 28 days, as suggested by our first hypothesis above.

Figure 4 shows the cumulative inter-arrival distributions of this non-bursty process compared to the actual data. Surprisingly, the distributions are almost identical. This suggests that in terms of the day-to-day activities of user browsing behavior, consumption of breaking news or other bursty topics simply does not play a significant part in the total volume of pageviews.

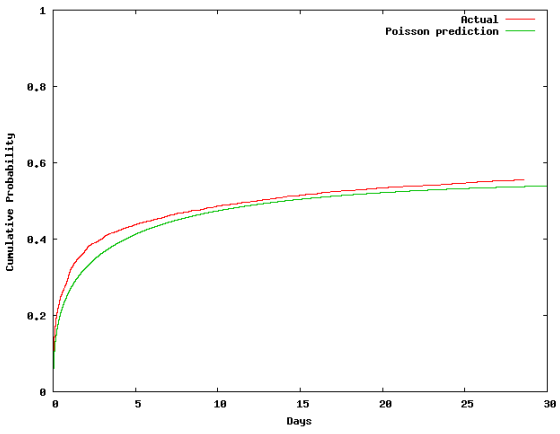

Figure 4: Poisson approximation of inter-arrival time distribution, compared to actual.

6. SEARCH BEHAVIOR

In this section, we give some insight into the nature of search sessions. There have been a number of detailed studies of the interactions with the search engine itself; see, for example [17]. However, there are fewer studies of search sessions that include behavior after the user departs the engine and begins to follow an information thread through the Web; [14] is an interesting work along these lines.

We define a search session as follows. Recall that a page is labeled MAIN SEARCH if it contains web search results from one of the top search engine — the pageview containing the initial search box is not labeled MAIN SEARCH. We consider the referral forest of a user, and study each referral tree in turn. We define the *search roots* of a referral tree to be the set of all nodes that are labeled as MAIN SEARCH, and that do not have any ancestor labeled MAIN SEARCH. A *search session* is then taken to be the set of pageviews in the subtree rooted at a search root. Notice that, by definition, search sessions may overlap with one another in time, but each pageview belongs to at most one search session. The mean number of search sessions per day per user is 0.57. The average number of pageviews in a search session is 6.3, and the average depth of the subtree rooted at a search root is 4.2. This implies that users do not simply perform a search and then explore the immediate results; they typically explore longer paths from the initial search results page.

We now consider the types of pages encountered by users in search sessions. By definition, the first page in a search session is of type MAIN SEARCH. We consider the distribution of types of other pages in the session. This should be viewed as the distribution of types of pages reached (directly or indirectly) from search. Table 13 gives this information. The makeup of these pages is seen to be quite different. SOCIAL and MAIL pageviews occur with an order of magnitude less frequency; search pageviews occur with significantly higher frequency, and other forms of search are marginally more frequent than in the general pageview distribution.

The aggregate picture one should take from this data is the following. On average, a search session takes around six pageviews. Of these, one is the initial search, on average there is one additional search, and the remaining four

Page type	% pageviews
UNKNOWN	48.4
SOCIAL	24.0
MAIL	9.5
MAIN SEARCH	5.9
MULTIMEDIA	3.3
PORTAL	2.6
OTHER SEARCH	2.1
HEAD LISTINGS	2.1
MULTIMEDIA SEARCH	1.8
ITEM SEARCH	0.3

Table 13: Pageviews with a search ancestor broken by page type.

Search session length	% pageviews containing search results
1	100.0
2	66.6
3	48.1
4	43.4
5	37.5
6	34.3
7	31.9
8	30.0
9	28.6
10	27.5
11	26.3
12–24	20–25
25+	15.3

Table 14: Fraction of search session pageviews spent at search engine, as a function of total pageviews in session.

pageviews are off-search. Table 14 gives a more detailed breakout of this picture. The table shows for each session length what fraction of time is spent on the search engine, versus browsing on the rest of the Web. As the table shows, longer sessions correspond strongly to more time spent engaging off the search engine with content on the rest of the Web.

7. CONCLUSIONS

We proposed a new *CCS taxonomy* of pageviews consisting of the following three high-level classes:

(1) Content (news, portals, games, verticals, multimedia) representing about half of all online pageviews;

(2) Communication (email, social networking, forums, blogs, chat) representing about one-third of all pageviews; and

(3) Search (web search, item search, multimedia search) representing about one-sixth of all pageviews.

We have presented a series of characterizations regarding the extent to which pages of certain types are revisited by the same user over time, and the mechanisms by which users move from page to page, within and across hosts, and within and across page types. We considered robust schemes for assigning responsibility for a pageviews to ancestors along the chain of referrals. We showed that mail, news, and social networking pageviews are insular in nature, appearing primarily in homogeneous sessions of one type. Search

pageviews, on the other hand, appear on the path to a disproportionate number of pageviews, but cannot be viewed as the principal mechanism by which those pageviews were reached.

Finally, we studied the burstiness of pageviews associated with a URL, and showed that by and large, online browsing behavior is not significantly affected by "breaking" material with non-uniform visit frequency.

Acknowledgments

We thank the Yahoo! editorial team for their help in labeling the webpages.

8. REFERENCES

[1] E. Adar, J. Teevan, and S. T. Dumais. Resonance on the web: Web dynamics and revisitation patterns. In *Proc. 27th CHI*, pages 1381–1390, 2009.

[2] E. Baykan, M. R. Henzinger, L. Marian, and I. Weber. Purely URL-based topic classification. In *Proc. 18th WWW*, pages 1109–1110, 2009.

[3] M. Bilenko and R. W. White. Mining the search trails of surfing crowds: Identifying relevant websites from user activity. In *Proc. 17th WWW*, pages 51–60, 2008.

[4] M. Bilenko, R. W. White, M. Richardson, and G. C. Murray. Talking the talk vs. walking the walk: Salience of information needs in querying vs. browsing. In *Proc. 31st SIGIR*, pages 705–706, 2008.

[5] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[6] A. G. Büchner, M. Baumgarten, S. S. Anand, M. D. Mulvenna, and J. G. Highes. User-driven navigation pattern discovery from internet data. In *Proc. WebKDD*, pages 74–91, 1999.

[7] R. E. Bucklin and C. Sismeiro. A model of web site browsing behavior estimated on clickstream data. *Journal of Marketing Research*, 11:249–267, 2003.

[8] I. V. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Model-based clustering and visualization of navigation patterns on a web site. *DMKD*, 7(4):399–424, 2003.

[9] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN Systems*, 27(6):1065–1073, 1995.

[10] O. Chappelle and Y. Zhang. A dynamic Bayesian network click model for web search ranking. In *Proc. 18th WWW*, pages 1–10, 2009.

[11] A. Cockburn and B. McKenzie. What do Web users do? An empirical analysis of Web use. *Intl. J. of Human-Computer Studies*, 54(6):903–922, 2001.

[12] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Probabilistic query expansion using query logs. In *Proc. 11th WWW*, pages 325–332, 2002.

[13] D. Downey, S. Dumais, and E. Horvitz. Models of searching and browsing: Languages, studies, and applications. *JASIST*, 58(6):862–871, 2007.

[14] D. Downey, S. Dumais, D. Liebling, and E. Horvitz. Understanding the relationship between searchers' queries and information goals. In *Proc. 17th CIKM*, pages 449–458, 2008.

[15] F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y.-M. Wang, and C. Faloutsos. Click chain model in web search. In *Proc. 18th WWW*, pages 11–20, 2009.

[16] E. Herder. Characterizations of user web revisit behavior. In *Proc. Workshop on Adaptivity and User Modeling in Interactive Systems*, 2005.

[17] 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*, 36:207–227, 2000.

[18] E. J. Johnson, W. M. Moe, P. S. Fader, S. Bellman, and G. L. Lohse. On the depth and dynamics of online search behavior. *Management Science*, 50(3):299–308, 2004.

[19] R. Jones and D. Fain. Query word deletion prediction. In *Proc. 26th SIGIR*, pages 435–436, 2003.

[20] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *Proc. 15th WWW*, pages 387–396, 2006.

[21] R. Kumar and A. Tomkins. A characterization of online search behavior. *IEEE Data Eng. Bull.*, 32(2):3–11, 2009.

[22] T. Lau and E. Horvitz. Patterns of search: Analyzing and modeling web query refinement. In *Proc. 7th UMAP*, pages 119–128, 1999.

[23] Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, and H. Li. Browserank: Letting web users vote for page importance. In *Proc. 31st SIGIR*, pages 451–458, 2008.

[24] P. Mayr. Website entries from a web log file perspective — a new log file measure. In *Proc. AoIR-ASIST Workshop on Web Science Research Methods*, 2004.

[25] Q. Mei, K. Klinkner, R. Kumar, and A. Tomkins. An analysis framework for search sequences. In *Proc. 18th CIKM*, 2009.

[26] A. L. Montgomery and C. Faloutsos. Identifying web browsing trends and patterns. *IEEE Computer*, 34(7):94–95, 2001.

[27] J. Morrison, P. Pirolli, and S. K. Card. A taxonomic analysis of what World Wide Web activities significantly impact people's decisions and actions. In *Proc. CHI*, pages 163–164, 2001.

[28] H. Obendorf, H. Weinreich, E. Herder, and M. Mayer. Web page revisitation revisited: Implications of a long-term click-stream study of browser usage. In *Proc. CHI*, pages 597–606, 2007.

[29] Y.-H. Park and P. S. Fader. Modeling browsing behavior at multiple websites. *Marketing Science*, 23(3):280–303, 2004.

[30] F. Radlinski and T. Joachims. Query chains: Learning to rank from implicit feedback. In *Proc. 11th KDD*, pages 239–248, 2005.

[31] A. Spink, M. Park, B. J. Jansen, and J. Pedersen. Multitasking during web search sessions. *Information Processing and Management*, 42(1):264–275, 2006.

[32] L. Tauscher and S. Grennberg. How people revisit web pages: Empirical findings and implications for the design of history systems. *Intl. J. of Human-Computer Studies*, 47(1):97–137, 1997.

[33] J. Teevan, E. Adar, R. Jones, and M. Potts. Information re-retrieval: Repeat queries in Yahoo's logs. In *Proc. 30th SIGIR*, pages 151–158, 2007.