Large Scale Log Analysis of Individuals' Domain Preferences in Web Search

Sarah K. Tyler[1], Jaime Teevan[2], Peter Bailey[3], Sebastian de la Chica[3] and Nikhil Dandekar[3]

[1]University of California, Santa Cruz
Santa Cruz, CA USA
skt@soe.ucsc.edu

[2]Microsoft Research
Redmond, WA USA
teevan@microsoft.com

[3]Bing, Microsoft Corporation
Bellevue, WA USA
pbailey@microsoft.com

ABSTRACT

Information on almost any given topic can be found on the Web, often accessible via many different websites. But even when the topical content is similar across websites, the websites can have different characteristics that appeal to different people. As a result, individuals can develop preferred websites to visit for certain topics. While it has long been speculated that such preferences exist, little is understood about how prevalent, clear, and stable these preferences actually are. We characterize website preference in search by looking at repeat domain use in two months of large-scale query and webpage visitation logs. We show that while people sometimes provide explicit cues in their queries to indicate their domain preferences, there is a significant opportunity to identify implicit preferences expressed via user behavior. Although domain preferences vary across users, within a user they are consistent and stable over time, even during events that typically disrupt normal search behavior. People's preferences do, however, vary given the topic of their search. We observe that people exhibit stronger domain preferences while searching than browsing, but that search-based preferences often extend to pages browsed to after the initial search result click. Since domain preferences are common for search and stable over time, the rich understanding of them that we present here will be valuable for personalizing search.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process

Keywords

User Preferences, Query Log Analysis, Web Search

1. INTRODUCTION

Web search users are not homogenous in their interests or tastes when selecting which search results to visit, any more than they are when selecting newspapers or magazines to read. Just as one person prefers to buy the New York Times while another prefers to buy the Wall Street Journal, so, too, do we observe that one person prefers to read movie reviews available on the Rotten Tomatoes website while another prefers them from IMDB's website. User preference for a website can be influenced by many factors, including the content, the way the content is communicated, perspective, technical detail, topical expertise level, and usability. Different individuals sometimes prefer sites that provide very different types of content that address the same information need. For example, someone may be more interested in reading Yelp reviews than visiting a company's Facebook fan page when searching for information about a local restaurant. Likewise, an individual with a strong political point of view may prefer to visit sites that support that point of view. However, individuals can have strong preferences even when the content from different sites is very similar. A Barnes & Noble rewards club member may prefer that online store to Amazon.com.

To better understand how website preferences are reflected in people's web search behavior, we present an analysis using large-scale query and browser logs of how often individuals turn to the same website for information when searching. We study people's explicit website preferences, where the site is named in the query, and their implicit preferences, which can be observed from click behavior. Previous work explored how individuals re-use search to re-find particular webpages [27][31], and looked at aggregate user patterns related to the revisitation of websites [2][17][26]. However, relatively little is known about how individuals use known websites during search. The work described in this paper extends previous investigations by exploring how individual users preferentially select domains. While repeat domain use may include the repeat URL visits studied in previous work, the behavior we study also includes instances when the user finds URLs from a repeat domain that they have not seen before.

While websites can be represented many ways, for the purpose of this analysis we represent them in terms of *domains* (short for "internet domain names"). Following a description of related work (§2) and the log data used (§3), we describe several ways to measure domain preferences that account for the overall popularity of the domain and capture both positive and negative preferences (§4). We use these approaches in §5 to characterize the scale and diversity of domain preference. We look at the stability of an individual's preferences over time and across different topics, and measure the differences in domain preference for search engine interactions versus general Web browsing. We find that people provide very strong explicit clues about their domain preferences in their queries, but that the implicit cues that can be identified via their behavior are more likely to benefit from search personalization. While domain preferences vary across users, an individual's domain preferences are stable over time, suggesting domain-based personalization is a promising approach. Additionally, while people exhibit stronger domain preferences while searching than browsing, search-based preferences can extend to pages browsed to after the initial search result click.

2. RELATED WORK

This work builds on recent research that has shown that domains can bias the perceived relevance of a search result. Ieong et al. [12] found that the same snippet text was judged to be more relevant when it was associated with a webpage from a well-known domain versus a lesser known domain. This aggregate bias was found to hold across many different queries and domains. We extend these findings to create a picture of individual domain preferences, showing that individuals sometimes prefer particular domains disproportionately from the rest of the population. To do this, we build on related research themes including Web log analysis, re-finding, and personalization of Web search.

Web Log Analysis: The work presented here uses behavior-based Web log analysis to explore domain preference in search and browsing. With the emergence of ubiquitous Web-based information, search engine query logs have become a rich source of data for increasing our understanding of user behavior when searching for and interacting with information. Silverstein et al. [22] published one of the earliest large-scale analyses of Web search, in which they presented metric-based summaries of various query, session, and user characteristics. Numerous insights into search patterns have been identified through analyzing large search engine query logs, including (among many others), interactions between search and browsing [5][23][34].

Because of their large scale, Web logs have provided a rich source to observe conditional (and not just aggregate) behavior, such as the effects of user demographics [33] and predictive models of user topical interests conditioned by different forms of context [2] or even specific user click behavior [19]. Information finding behavior conditioned by an individual's prior search activities is particularly relevant to this paper. We discuss the valuable insights that search logs have provided in more detail below.

Information Re-Finding: The re-finding of previously viewed URLs is one way that interaction behavior conditioned on prior activities that has been studied. Teevan et al. [27] analyzed the queries issued by 114 anonymous users in one year's data from Yahoo search logs, and observed that users re-found previously found URLs for 40% of all queries. In a larger study involving tens of millions of users, Tyler and Teevan [31] further examined re-finding using both search and browser logs from Microsoft's Bing search engine and Windows Live Toolbar. They observed distinct query patterns and behaviors relating to re-finding that differ from other search patterns, and explored differences between same-session and cross-session re-finding. Shokouhi et al. [21] analyzed repeated result use within a session, and showed that users engage differently with repeats than with results shown for the first time. Our study builds on previous work to characterize how people re-find not just previously viewed URLs, but also previously viewed domains. We study the re-finding of individual URLs as a subset of domain re-finding behavior.

Although domain preference has not been well studied in the context of Web search, researchers have explored the consistency of people's domain use in the context of Web browsing. Tauscher and Greenberg [26] found that the majority of webpage revisits occurred within a small set of websites. Obendorf et al. [17] found significant influence of site type on page revisitation, and that the frequency at which a site was visited correlated with the variation of pages within the sites that were visited. Adar et al. [2] characterized webpage revisitation signatures via fast, medium, slow and hybrid revisitation curves, and found that fast revisitations often occurred clustered within a particular website.

Search Personalization: The study of Web-based information finding behavior as conditioned by an individual's prior search activities is interesting in part because it can be applied to the personalization of Web search [18]. Web search personalization has been an increasingly active area of investigation in recent years. Many different types of behavioral data have been used to represent an individual's interests, including the searcher's query history [20][21][24], browsing history [16][25], and rich client-side interactions [4][6][16][28]. The time horizon of the behavioral data used can range from short term, context based, to longer-term, interest based profiles. For example, some researchers have explored queries and clicks from the current session [20][21], while others have included information from months or years ago in their user profile [28]. Dou et al. [8] showed that both long- and short-term profiles are important for personalized search performance.

Of particular relevance to this paper, there is some evidence that previously visited domains can be useful for Web search personalization. Teevan et al. [28] found that boosting Web search results with URLs from domains that the user had visited in the past yielded significant improvements in result ranking for 131 manually labeled queries. Matthijs et al. [15] confirmed these findings via an in-field user study on personalized re-ranking. The personalization of re-finding activities has also been investigated from various standpoints, including predicting personal navigation in a Web search environment [29], email re-finding [9], and using re-finding signals for personalization [32]. Personalization for re-finding queries has been successful because it is limited to queries for which there is very strong signal [29]. We find that the signal is also very strong for domain-based personalization, but unlike approaches limited to re-finding, domain-based personalization can be applied to queries targeting new content.

3. DATASETS

Most of the analysis presented in this paper is based on search log data collected from Bing, a leading Web search engine. We supplement these findings with analysis of browser logs collected via a popular internet toolbar in order to understand how domain preference impacts behavior after visiting a search result and how search preference compares with general browse preference.

3.1 Search Log Dataset

The search log data we analyzed is a sample of query logs from Bing, dating from November 1, 2010 to December 31, 2010. The sample includes the queries issued by 230 million users, as well as the URLs of the top 10 results returned for those queries and the search results people clicked. Seventeen million users were deemed *Frequent Users*, meaning they issued at least 25 queries during the study period. Since understanding people's patterns of search behavior over time requires multiple observations from the same user, all of the analysis in this paper focuses on these Frequent Users.

The search results presented to Frequent Users came from 26.1 million unique domains, with an average 1.32 unique pages per domain. Of the unique domains, 1.49 million were *Frequent Domains*, meaning they appeared in the search results for at least 10 times different queries from a single user, regardless of whether they were clicked or not. The same domain is much more likely to occur in multiple result sets than the same specific URL is; as a comparison, only 190 thousand unique pages occurred in the search results for at least 10 queries for a single user.

Of the search result clicks, half (49.7%) were on results from domains that had previously been found by the same user via search (i.e., where the user previously clicked on a search result for the same domain). We refer to a search result click on a domain that has already been found by an individual as a *Repeat Domain Click* (see Table 1). Approximately two thirds (69.0%) of the *Repeat Domain Clicks* were clicks on exactly the same URL that the user had previously clicked, with the remaining third (31.0%) on results the user had not clicked previously.

3.2 Toolbar Dataset

We supplement the search logs with toolbar data to see how users interact with preferred sites after searching for them, and to compare domain preference when searching with browsing. The toolbar logged the webpage visits of opt-in users. For consistency, we use browser behavior for the same time interval (November 1

Category	Term	Definition
Users	*Frequent User*	A search engine user who issued at least 25 queries during the two months sampled in the logs.
Domain	*Domain*	The level at which the domain was available for public registration. Usually the second level domain (e.g., microsoft.com, amazon.com), but sometimes includes a country code (e.g., amazon.co.uk).
	Frequent Domain	A Domain that appeared at least 10 times in the search results for a single user.
Domain Preference	*Explicit*	A user directly signals a preference for a particular domain by explicitly mentioning it in the query and then clicking on a result from that domain.
	Implicit	A user's preference for a particular domain is observed based on the domains the user visits.
Query	*Domain is Query*	A query where the query itself is the domain name of the search result click, with or without top level domain information like ".com" or ".org". Whitespace is ignored, and terms must appear in order so *blog spot* matches blogspot.com whereas *spot blog* does not.
	Domain in Query	A query in which the user mentions the specific domain (e.g., *"andrea dailey" blogspot*) of the result they click. Punctuation, case, and URL identifiers are ignored, so *sewing needles joann's* is understood to refer to the domain joann.com, and *domain info wikipedia.com* to refer to wikipedia.org.
	Navigational (Nav) Query	A query targeted at a particular webpage. A query instance is considered navigational if (1) across all users the query has a high click-through rate on a single URL and low click entropy, and (2) the user issuing the query clicks on a result from that domain. E.g., most people click http://www.wsdm-conference.org/2014 after the query *wsdm*. Users who do have issued a *Nav Query*.
Search Result Click	*Repeat Domain Click*	A search result click where the user clicks on a URL from the same domain the user has clicked on during a previous search.
	Root Level Domain Click	A search result click on the top level page for a domain, rather than a subpage within the domain. The subdomain "www." may or may not be present. For example, http://www.ebay.com is considered a top level page for ebay.com, but http://my.ebay.com and http://www.ebay.com/login.php are not.
	Repeat URL Click	A search result click where the user clicks on the same URL, not the root level, that they clicked on during a previous search.
	New URL Click in a Repeat Domain	A search result click that is not a *Repeat URL Click* or a *Root level Domain Click* but that is a *Repeat Domain Click*. For example, a user who searches for many L.A. Times articles may visit a new article.
Trail	*Click Trail*	A series of URLs visited while clicking on links within the pages, or by navigating via the back button to an earlier URL in the trail. In keeping with prior work [34], we consider a trail to end when the user enters an address in the address bar, submits form data, logs into a website, or visits a bookmark.
	Search Result Trail	A click trail where the initial page is a search result. A search trail can also end when the user returns to the search engine.
	Browser Trail	A click trail where the initial page is reached by the user typing an address in the address bar.

Table 1. Important domain-related definitions used throughout the paper.

to December 31). Note, however, that the toolbar users may not be exactly the same as the users in the search logs. Therefore, we extracted user query and click data across major search engines (including Google, Yahoo and Bing) from the toolbar logs in order to relate browsing and searching behavior for the same user. We use the larger, richer search engine query logs for most of our analysis, except when a combination of searching and browsing is necessary, but note that the search behavior observed in both the search and toolbar logs is generally consistent.

The toolbar dataset contains the browsing behavior of 12.3 million users, who issued 84 million queries with clicks and typed an address into the address bar 286 million times. We filtered this to the 3 million users who had 25 or more *Browser Trails*, which are trails of clicks initiated by typing in an address (see Table 1). The filtered logs contain 223 million *Browser Trails* with an average length of 5.92. Searching was less common than browsing, with only 669 thousand users issuing 25 or more searches.

4. DEFINING DOMAIN PREFERENCE

We now describe how we measure domain preference, and define the terms used throughout the paper. A summary of the definitions can be found in Table 1. A user's preference for a particular domain can be either *Explicit,* with the user directly signally a preference, or *Implicit*, with the preference observed via the user's behavior. Both are discussed in greater detail below.

4.1 Explicit Preference

A user can directly signal a preference for a particular domain by explicitly mentioning it in the query. We identify three types of queries that typically indicate an explicit domain preference (Table 1, *Query* section). The user can mention a domain in the query text (*Domain in Query*), issue a query where the query text is itself the name of a domain (*Domain is Query*), or issue a navigational query (*Navigational*).

Navigational queries are queries where users consistently click on the same URL and nothing else. While the other types of explicit domain preference queries are identified using the query text, navigational queries are identified via behavioral data. We nonetheless consider navigational queries to be an example of *Explicit* preference for a particular domain because the query text for these queries is usually the name of an organization associated with the clicked website (e.g., *wsdm*), or a specific key word (e.g., *windows update*) that is only used to reference one website. Since these queries correspond to clicks on a single URL, we assume the user's intention when issuing the query was to navigate to that particular URL.

While *Explicit* domain preference is signaled in the user's query, such a preference may not extend beyond the query. In the case of the domain http://facebok.com, for example, the user may have no further interest in the domain outside of an *Explicit* preference query. For this reason, we also explore users' implicit domain preferences over their entire search history.

4.2 Implicit Preference

A user's implicit preferences for a particular domain can be observed from the user's click history. A user is considered to have an *Implicit* preference for a domain if, when given the option, that user chooses to click on search results from that domain more than expected. There are multiple ways an individual can signal a domain preference via their clicks, summarized in the *Search Result Click* section in Table 1. Most generally, a person can click on a URL from the same domain as a URL the user has previously clicked on. We call these *Repeat Domain Clicks.* In the case of re-finding, the *Repeat Domain Click* is on a URL that the user has clicked previously, which we call a *Repeat URL Click.* Users may also click on new pages from a previously visited domain, and we refer to this as *New URL Clicks in a Repeat Domain*. Such behavior can indicate a more general preference for the domain beyond a preference for specific pages. Because the behavior surrounding top level page for a domain (e.g., http://www.ebay.com) can be unique, we separate out these clicks and refer to them as *Root Level Domain Clicks*.

The presence of a domain in a user's click history does not inherently indicate a preference, as some websites are more prevalent than others. Clicks on webpages from popular domains like Wikipedia appear frequently in people's search histories in part because the domain occurs frequently in search results. If a user clicks on results from a domain less frequently than might be expected given its popularity, the user probably does not have a preference for it. To identify a user's implicit domain preference, we therefore rely not only on how frequently the user interacts with results from a particular domain, but also how frequently all users interact with results from that domain.

Since the task of identifying domains that are uniquely interesting to an individual can be framed similarly to the task of identifying terms that are uniquely interesting in a document, we measure the strength of domain preference by modifying two standard information retrieval techniques: TF.IDF and KL Divergence. To do this, we consider the domain of a search result a user clicks on as analogous to a term. A user's search history is viewed as a collection of terms, and a user as a document.

TF.IDF can be used to find terms that are disproportionately more frequent in a document than in the overall corpus. We find the relative weight of a domain in the user's search history using TF.IDF Pref (Equation 1) where *d* is the domain, $\{d: d \in u\}$ indicates the user *u* has clicked on the domain, U is the set of all users, and tf(*d*) is the frequency that URLs with domain *d* are clicked in the search results.

$$TF.IDF\ Pref(d) = tf(d) * Log\left(\frac{|U|}{|\{d: d \in u\}|}\right) \qquad (1)$$

Positive TF Preference We consider a domain to have a Positive TF Preference for a user if the domain has a TF.IDF Pref score greater than the median TF.IDF score for that user.

KL Divergence is used to compare the distribution of terms in two documents. It can also be used to compare the distribution of domains in two search histories. The probability of a domain click is the smoothed percentage of clicks on the domain relative to all clicks in the user's search result history. KL Divergence can then be used to measure how much the user's history diverges from global distribution of search result domain clicks, where the global probability is derived from the smoothed user probabilities.

We consider the relative weight of a single domain in the distribution using Equation 2, where P_u and P_g are the smoothed probabilities of a domain click from the user and global distributions, respectively. This equation is sometimes referred to as pointwise KL Divergence in the literature [30] and has been shown to be effective for identifying phrases that distinguish between documents in terms of political blogs [1][11], blog mining for market intelligence [10] and identifying different events in temporal queries [13].

$$KL\ Pref(w) = P_U(d) * Log\left(\frac{P_U(d)}{P_G(d)}\right) \qquad (2)$$

Positive KL Preference We consider a domain to have a Positive KL Preference if the KLPref score of the domain for a user is greater than the median of the non-negative KLPref scores for that user.

If a user clicks on the domain more than average, that user is likely to have a positive preference. A user can also exhibit a negative preference in a domain by clicking on it less frequently than average. Since KLPref can be negative if the probability of the user to click on a domain is less than the global probability of a click, we can use it to identify negative preferences.

Negative KL Preference We consider a domain to have a Negative KL Preference if it has a score less than the median non-positive KL score for the user.

When a result from a particular domain is displayed and not clicked, this does not necessarily indicate a negative preference. A domain may only appear in the results rarely, or may be ranked so low that it is not careful inspected by the user. To avoid passing judgment on a domain with sparse data, we only consider the implicit preferences of *Frequent Users* (i.e., users with at least 25 clicked queries) for *Frequent Domains* (i.e., domains that have occurred at least ten times within the user's search history).

For the majority of our analysis we compute the KL Preference and TF.IDF Preference scores across the user's entire search history. Preference for a domain can differ depending on the type of content the user is seeking. Therefore, we also consider a category based KL Preference and TF.IDF Preference in §5.2.1.

5. FINDINGS

Using these measures of *Explicit* and *Implicit* domain preference, we analyze the search logs to understand the prevalence and stability of people's preferences. We also use the toolbar logs to compare people's preferences across searching and browsing, and to see how they interact with a domain targeted during a search.

5.1 Prevalence of Preference

Our analysis suggests that domain preferences is very common. *Repeat Domain Clicks*, which are clicks on URLs from domains previously visited via search, account for 49.7% of all search result clicks. To build a picture of these clicks, we first look at explicit domain preference, where the user specifies a domain in their query, and then explore implicit preference, where the user clicks a domain more than expected.

		Repeat Domain Clicks		
		Root Level Domain Click	Repeat URL Click	New URL Click in a Repeat Domain
Expl. Pref.	Domain is Query	34.35%	2.57%	0.56%
	Domain in Query	1.40%	3.15%	1.98%
	Nav Query	3.08%	2.26%	0.41%
	Other Query	8.42%	14.72%	27.09%

Table 2. Break down of the types of queries used in searches resulting in *Repeat Domain Clicks*.

5.1.1 Explicit Preference

Explicit domain preference was often expressed in the user's query prior to a *Repeat Domain Clicks*. In total, almost half of all *Repeat Domain Clicks* were preceded by a query that indicated an explicit preference. Table 2 breaks down the occurrence of explicit preference by the different ways queries were used to express it. For 44.02% of all *Repeat Domain Clicks*, the domain that the user clicked appeared directly in the user's query, either in the query (*Domain in Query*) or as the entire query (*Domain is Query*). For an additional 5.75% of the *Repeat Domain Clicks*, the click was preceded by a *Navigational Query*.

The clicks following a query with explicit preference are further broken down by the type of repeat click that occurred, including *Root Level Domain Clicks,* re-finding clicks (*Repeat URL Click*), and clicks on a new URL from a repeat domain (*New URL Click in a Repeat Domain*). Our data suggest it may be particularly easy for users to explicitly signal domain preferences for pages which are already familiar. For all three types of explicit queries, users were more likely to click on the root level domain page or a repeat URL than a new URL. In contrast, 30.04% of *Repeat Domain Clicks* in general (or 14.92% of all search result clicks) were on new URLs (*New URL Click in a Repeat Domain*). The majority of such clicks were not preceded by an explicit domain preference.

Explicit preferences are common and easy to detect, and appear to be well-supported by search engines. The most common form of explicit preference, *Domain is Query*, tended to lead to a *Root Level Domain Click*, and in most of these cases the root level domain was listed in the top rank position. Additionally, *Nav Queries*, by definition, tend to lead to a specific URL click that is easy to identify. Since these two cases account for 80.57% of all explicit preference queries, most explicit preference queries can be accommodated by either returning a specific URL (in the case of a *Nav Query*) or root domain (in the case of *Domain is Query*).

While explicit preference queries are already well supported, it may be possible to use them to improve other, less well supported searches with domain preference. In §5.1.3 we will show that the signals we observe relating to explicit and implicit preferences sometimes overlap. Explicit preference queries could be used to identify preferred domains for queries with implicit preferences, especially when the user ultimately clicks a new URL from a repeat domain.

5.1.2 Implicit Preference

Like explicit preference, implicit domain preference was common in the logs, with 94.56% of all *Frequent Users* exhibiting an implicit positive preference for at least one *Frequent Domain* in their search history. Individual users, however, displayed an implicit preference for only a handful of domains. The histogram of the number of preferred domains per user is displayed in Figure 1. The average (median) user had one positive domain preference according to both TF and KL preferences, and only 1.8% of users had more than 10 implicit domain preferences.

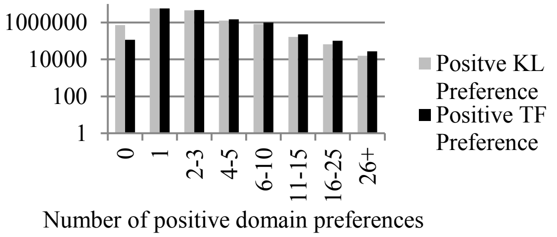

Figure 1: Histogram of the number of frequent user with positive domain preferences for Frequent Domains. Most users prefer only a handful of domains.

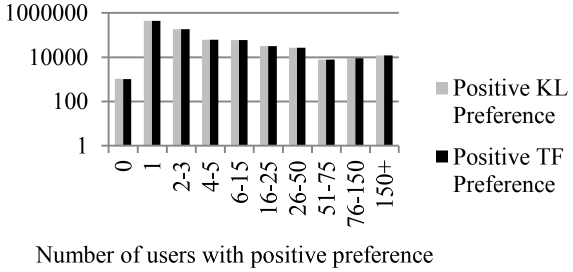

Figure 2: Histogram of the number of Frequent Domains that are positively preferred by Frequent Users. Most domains are only preferred by a handful of users.

The implicit domain preferences we observed were highly individualized, and most users' preferences were unique. A histogram of the number of users who preferred a given domain is shown in Figure 2. A domain had a mean average of 38.4 users with a positive KL Preference for it, and 44.0 users with positive TF Preference for it. Only twelve thousand domains (roughly 1.47% of the frequently occuring domains) were preferred by 150 users or more.

Users were more likely to have a positive implicit preference than a negative one. This is in part because a single click could account for as much as 4% of a *Frequent User*'s click history, which, even when smoothed, is larger than the average percentage of clicks most domains receive across all users. As a result, most domains that received a click registered as positively preferred, and it was rare for a negatively preferred domain to show up as having received a click but fewer than expected. Further, while there are typically many domains in a user's search history that are not clicked, individual unclicked domains tend to occur infrequently. The most frequent domains in the user's search history tended to be clicked at least once, and thus were more likely to register as positively preferred than negatively preferred. To avoid data sparsity, all domains included in our analysis were required to occur in a user's history at least 10 times. Even when this restriction was relaxed to just 5, only 4.37% of users had one or more domains that they appeared to dislike based on their implicit interactions.

The most common negatively preferred domains are shown in Table 3, along with percentage of users who, given a preference for the domain, preferred the domain positively or negatively. These domains are all popular Web destinations, and, as can be seen in the table, it was much more common for people to have a

Category	Domain 1 (D1)	Domain 2 (D2)	$P(Pref_{D1}\|Pref_{D2})$	$P(Pref_{D1})$	$P(Pref_{D2}\|Pref_{D1})$	$P(Pref_{D2})$
Social	Facebook	Twitter	60.95%	20.70%	0.03%	0.01%
	Facebook	MySpace	31.30%	20.70%	18.61%	11.41%
Shopping	Etsy	Ebay	0.05%	2.03%	0.42%	18.60%
Travel	Expedia	Orbitz	21.60%	2.23%	10.84%	1.12%
Books	Barnes & Noble	Borders	27.73%	3.70%	5.80%	0.77%
Movies	Rotten Tomatoes	IMDB	3.51%	1.00%	48.36%	13.79%
News	Fox News	MSNBC, Huffington Post	10.15%	3.98%	17.56%	6.89%
Recipes	Cooks.com	Allrecipes.com	38.65%	7.57%	60.67%	11.88%
Reference	Wikipedia	eHow	55.46%	29.18%	7.57%	3.98%
Technology	Microsoft	Apple	22.54%	0.40%	2.20%	0.04%
TV	Hulu	Netflix	10.17%	2.99%	10.20%	3.00%

Table 4: Implicit domain preference in pairs of similar domains. $P(Pref_D)$ is the probability a user will have an implicit preference in the domain D.

	% of users given preference for domain		
	+ KL Pref (10 instances)	- KL Pref (5 instances)	- KL Pref (10 instances)
Facebook	72.9%	30.7%	5.4%
Yahoo	49.9%	27.2%	12.9%
Wikipedia	51.9%	17.6%	8.6%
Google	66.3%	17.4%	5.0%
YouTube	73.0%	10.5%	3.1%
Craigslist	88.9%	6.2%	0.4%
Amazon	77.6%	3.6%	0.7%
Answers.com	78.8%	2.4%	0.9%

Table 3. Examples of the most negative preferred domains that are frequently occurring in the user's search history.

positive preference for these domains than a negative one. Their popularity may be one reason why these domains tend to rank highly for multiple queries. However, our data suggest that even though the domains are popular for many users, they are not well suited for all users.

In many cases a preference for one domain indicated an increased preference for another domain. We can illustrate this this by considering pairs of related domains that offer similar content. Table 4 shows the conditional probability that a user will have an implicit preference for one domain (e.g., Facebook) given an implicit preference for a similar domain (e.g., Twitter). The data suggests that general interests in a topic outweighs any specific within-topic domain preferences. For example, we see that a user who preferred search results from Facebook was also more likely to prefer results from MySpace, and a user who preferred cooks.com was more likely to also prefer allrecipes.com. Even for domains such as Fox News and MSNBC, where individuals are thought to prefer one and not the other, there is, nonetheless, a positive correlation in preference.

While explicit domain preference appears well supported by search engines, implicit domain preference seems to be less well supported. The average rank of results clicks for implicit preferences was 1.61, much lower than the average rank of 0.33 for explicit preferences. The strong implicit domain preference we observe suggests a new opportunity for search result personalization. User profiles incorporating KL Divergence-based features to capture individual variation have been shown to be predictive of personalized search result selection in recent studies. Examples include Bennett et al.'s study on using inferred location preferences to personalize search [3], and Kim et al.'s study on reading level and topic influences on user search behavior [14].

Probability of Implicit preference	Implicit pref. measure	
	TF IDF	KL Divergence
P(Pref)	58.91%	55.83%
P(Pref | Domain is Query)	61.32%	61.32%
P(Pref | Domain in Query)	59.93%	56.07%
P(Pref | Navigational Query)	60.33%	63.56%
P(Pref | Root Level Domain Click)	65.52%	63.56%

Table 5. Probability of implicit preference given explicit preference. Explicit preference does increase the probability of implicit preference, but less than *Root Level Domain Click.*

5.1.3 *Overlap in How Preference Is Expressed*

Domains that are explicitly preferred are likely to also be implicitly preferred. Table 5 shows the conditional probabilities of implicit preference given explicit preference. While an explicit preference for a domain indicates a greater likelihood that the user will also have an implicit preference, the increase in likelihood may be better explained by other factors. For example, *Root Level Domain Clicks* are often stronger indicators of implicit preference than explicit preference queries. As noted earlier, explicit preference queries *Domain is Query* and *Nav Query* were more likely to lead to a *Root Level Domain Click*. As a result, the conditional probability of implicit preference given explicit queries may be influenced by this click action.

If a user clicks on a domain and not a subpage, it may be an indication that the whole site and not just a URL is relevant to the user's information need. It may also be the case the user is familiar with the domain, and willing to search within the domain or navigate within the domain. For users that are not *Frequent Users* and thus do not have rich history data to use for domain-based personalization, it may be possible to use their *Root Level Domain Clicks* to predict implicit domain preferences.

5.2 Stability of Domain Preference

We now look at the stability of an individual's domain preferences across different topics and over time. For example, a

Click Δ from general	% with a pref. in >1 category	User pref. across all categories			Pairs w/ pref order flipped
		Positive	Negative	Changes	
All	14.56%	20.19%	26.17%	53.64%	13.46%
≥ 0.1%	10.85%	26.89%	27.72%	45.38%	12.27%
≥ 1%	7.52%	34.71%	22.75%	42.55%	7.54%
≥ 5%	2.46%	71.95%	2.85%	25.20%	3.32%

Table 6. The percentage of domains where individual users diverge beyond some threshold from the general population. The middle columns show the type of preference (always positive, negative, or both) a user had for a domain across categories. The right column shows the percentage of domain pairs where the preference order flipped (i.e., where one domain was preferred above the other in Category 1, and vice versa in Category 2).

user may turn to the New York Times for political commentary and the LA Times for entertainment news, even though both sites cover similar content. A user's preference for content from a particular domain could also change over time as the user discovers new and interesting websites, or as the user's interests evolve. We show that domain preference is somewhat sensitive to topic, but stable over time.

5.2.1 Domain Preference across Topics

In order to explore how the topic of a search influences implicit domain preference, we look at whether the same user exhibited similar domain preferences in different categories. To do this, we classified webpages using a language-based topic classifier trained on the top two levels of the Open Directory Project (ODP, http://www.dmoz.org) using the approach described by Collins-Thompson and Bennett [7]. The URLs of second level topics were crawled to create training data for the classifier. The classifier was then used to predict one or more categories for the URLs visited in our search log. Not all URLs in the search results were given a category for a variety of reasons, including the fact that some pages had not text or no longer existed at the time of analysis. When we were unable to classify a URL that a user visited, we defaulted to classifying the text of the domain page. All of the categories of all of the URLs that a user visited were aggregated to create a user-specific set of categories per domain. Broad websites, such as Wikipedia, could be labeled with only a handful of categories for a user if the user only visited the site for a specific set of topics.

Up until now our analysis has focused on domains that appear frequently (at least 10 times) in the user's search history. However, only 5.7% of the domains in people's histories rose to the level of being *Frequent Domains*. Since even fewer of these appeared in multiple categories for the same user, we consider all domains in this analysis and not just *Frequent Domains*. Likewise, we consider any domain with a positive KL preference score to be positively preferred, and any with a negative KL preference score to be negatively preferred. We now look more closely at the 14.56% of the domains in our users' histories that were classified into multiple categories to determine whether the sign and magnitude of their preference was consistent across category. The column labeled "Changes" in Table 6 shows the percentage of instances where the implicit preference for a domain was positive in one category, and negative in another (i.e., the domain experienced a preference flip). About half (53.64%) of the domains that were present in multiple categories (roughly 7.8% of domains in the user's search history) experienced a preference flip.

Some of these preference flips may be small in magnitude. Even a *Frequent Users* may only have a small number of domains in any given category, which could create a precision error when comparing the user's percentage of clicks on the domain to the global percentage of clicks. A domain may be only slightly negatively preferred in one category, and only slightly positively preferred in another. For this reason, we calculated the percentage of preference flips where the user's percentage of search result clicks on a domain differed from the global percentage of clicks by varying thresholds (see Table 6). With a threshold of 5%, only 2.46% of domains are present in multiple categories, and 25.2% of these domains (less than 1% of the domains in the users search history) experienced a flip. In other words, for most users and domains, the user is only likely to have a disproportionate click rate on the domain for a single category.

In addition to direction (positive or negative), the magnitude of the preference can also change across categories. For example, a user may prefer Rotten Tomatoes to IMDB for science fiction movie reviews, and IMDB over Rotten Tomatoes for romantic comedy movie reviews, yet still have a positive preference for both domains in both categories. We consider such domains where the order of preferred domains differs per category to be pairwise preference flips. As shown in the last column of Table 6, only 13.46% percent of all domain pairs have their preference ordering flipped between categories. Compared to 53% of the domains that have the sign of their preference change, most preference changes do not affect the order of the preferred domains, and thus would not affect a preference based re-ranking of search results.

5.2.2 Domain Preference over Time

The previous section showed that an individual's implicit domain preference can sometimes vary across topic. We now look at how it changes over time. To do this we compare the KL Divergence of the distribution of domain clicks for a user on given day to an immediately preceding time interval in the user's search history. KL Divergence measures the similarity between two distributions: lower scores indicate the distributions are more similar while higher scores indicate they are more dissimilar. The per-user KL Divergences are averaged across *Frequent Users* and plotted in Figure 3. All click counts are smoothed by 0.25 clicks. We note that the difference between the average KL Divergence score for a given day and the average score for two days in the future is statistically significant ($p < 0.005$) for all but one data point (December 21st with 2 weeks of history.) The difference between average KL Divergence scores on consecutive days, however, is not always statistically significant as the KL Divergence curves tend to decrease slowly over the work week.

In general, click preferences appear to exhibit a weekly pattern, though the peaks differ for different history lengths. The KL Divergence curves using several weeks of history have a slightly lower average KL Divergence just before the weekends, while the KL Divergence curves by session and day yield a slightly lower average KL Divergence on the weekend. This shows that weekend behavior tends to be more consistent than weekday behavior, and that weekend and weekday click distributions are different.

Our logs consist of search behavior from users in the United States and cover several US holidays, including Thanksgiving (the fourth Thursday of November) and Christmas (December 25). While not all users would have celebrated these holidays, many may have an altered work schedule and engaged in atypical activities, such as shopping and visiting family. Analysis of US logs typically shows atypical search behavior during the holidays.

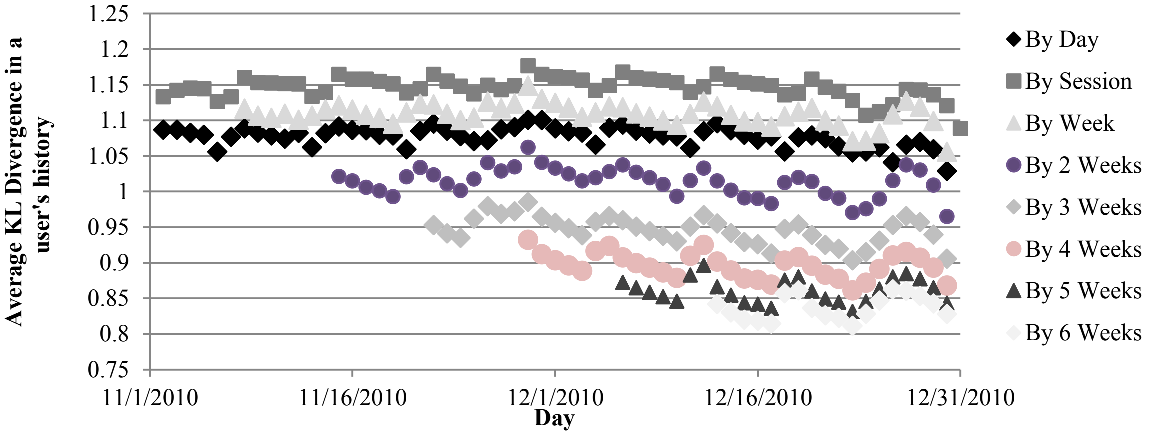

Figure 3: The KL Divergence between the distribution of domain clicks in the current time period to the previous time period for each user, averaged. Follows a daily pattern with the weekend being the most consistent (lowest KL Divergence).

Thus we expected to observe a shift in user interests around this time period. We grouped the first fifteen days of December as *Non-Holiday*, and the last fifteen days as *Holiday*. While there were 20% fewer user data points in the *Holiday* group than in the *Non-Holiday* group, indicating users were less active during the holidays, the average KL Divergence for the *Holiday* period was 1.1% to 1.6% lower than the *Non-Holiday* period when using less than 5 weeks of data, 0.5% less for 5 weeks of data, and only 0.5% more for six weeks of data. Thus, users were typically more consistent with themselves over the holiday portion of the month when using less than six weeks of history data.

5.3 Preference When Searching vs. Browsing

It is possible that the strong consistent domain preferences we observe when people search exist because people have strong domain preferences when using the web in general. However, by comparing how domain preference in search differs from domain preference in browsing, we find that there is a stronger pattern of preference while searching than in browsing.

We study search and brows behavior by comparing the click entropy between a user's *Search Result Trails* and *Browser Trails*. As shown in the Table 1, a *Click Trail* represents the series of URLs that a user visits while clicking on webpage links or the browser back button. *Search Results Trails* are the trails that begin with a search result page, while *Browser Trails* are the trails were the initial page was reached by typing the address into the address bar. Entropy is a measure of uncertainty, and we use click entropy to measure the distribution of domains clicked on a trail. The larger the click entropy, the more random the click distribution is. The entropy of a user's *Search Result Trails* can be compared with entropy of their *Browsing Trails*.

However, the different characteristics of *Browser* and *Search Result Trails* impact our ability to do this without first controlling for several factors. As noted in §3.2, our logs contain many more *Browser Trails* than *Search Result Trails*. To avoid sampling issues, we only consider users with more than 25 trails of each type. *Browser Trails* are typically longer than *Search Result Trails*, and longer trails are more likely to contain multiple instances of a domain and have lower entropy than short trails. This is because webpages tend to have more URLs pointing to internal domains than new external domains. To address this, we look at the entropy of two types of domain distributions: across all trail clicks and across trails.

To calculate the entropy across all trail clicks, we measure the probability of a click as a straightforward count of domains visited over all domains in all *Search Result Trails* or all *Browser Trails*. In this method, longer trails will contribute to there being lower entropy than shorter trails do. To calculate the entropy across trails, we consider the frequency of domain visits to be the number of trails in which the user visited the domain. The corresponding probability of a domain is the number of trails with the domain, over the number of (domain, trail) tuples. In this method, longer trails will likely contribute to a higher entropy since they have more potential for multiple domains. We find users were likely to have more entropy in their *Browser Trails* than their *Search Result Trails* using both metrics. Across all clicks, the entropy was 69.4% higher for *Browser Trails*, and across trails it was 78.4% higher. This means that people were more likely to stay within known domains when searching than when browsing.

The fact that people are more consistent when searching seems counter-intuitive because Web search is typically thought of as the seeking of new information. When a user clicks on a search result (which is the first step in a *Search Result Trail*) that click could be on a completely new domain to the user. On the other hand, when a user enters an address in the search bar to start a *Browser Trail,* that address is probably a known URL that the user has previously visited. To explore this, we further broke the *Search Result Trail* entropy down into three component parts: the start (the search result click), the end (the last click in the trail), and the middle (all remaining clicks). The entropy for the final click was higher than the entropy for the first search result click for 77.41% of all users. Thus it appears that people were actually more consistent in the results they chose, with greater uncertainty existing in where they would end up than where they would start.

There are multiple possibilities as to why there is less entropy in the initial search result click than in the final destination of a *Search Result Trail*. The user could be using the search engine to find a trusted place to begin exploration. For example, a user might access a Wikipedia article entry via a query and then explore the links for additional information. Another explanation is that it is easier to formulate a query on a known topic (or even issue the same query [27]), which in turn leads to finding previously visited websites, while navigating within a page can lead to the discovery of new topics and ideas. It also could reflect

		Staying power of search result domain within a Search Result Trail		Returning to search result domain given visits to other domains in the trail	
		Num. hops during initial domain visit	Prob. of visiting another domain	Prob. of returning to initial domain	Num. hops before returning to domain
Explicit Preference	Not Domain is Query	2.86	10.16%	3.15%	4.75
	Domain is Query	3.15	7.56%	5.34%	4.86
	Not Domain in Query	2.86	11.14%	2.96%	4.66
	Domain in Query	2.92	7.59%	4.16%	4.98
	Not Navigational Query	2.82	11.47%	2.72%	4.63
	Navigational Query	2.98	7.28%	4.80%	4.97
Implicit Preference	- KL Preference	2.60	11.42%	1.35%	4.48
	No KL Preference	2.91	10.68%	3.22%	4.60
	+ KL Preference	2.86	9.47%	3.30%	4.90
	No TF IDF Preference	2.91	10.90%	3.19%	4.62
	+ TF IDF Preference	2.85	9.39%	3.29%	4.87
	No previous visits	2.83	8.15%	3.06%	4.62
	Infrequent visits	2.93	9.09%	3.71%	4.88
	Frequent visits	3.06	10.64%	3.70%	5.32

Table 7: User interaction with preferred domains within a *Search Result Trail.*

the generalized domain bias towards more popular or trusted sites reported by Ieong et al [12], or search engines may be surfacing these same popular websites, reducing the overall entropy in the domains of the search result clicks.

5.4 Users Interaction with Targeted Domain

We now look more closely at people's *Search Result Trails*, broken down by whether they begin with a search result click on a preferred domain or not. We find that users interact differently with the pages from preferred domains, taking more hops within the preferred domain and being more likely to stay within it.

Table 7 summarizes how likely a user is to stay within the domain of a clicked search result, and how likely the user is to return to that domain if they leave. Statistics are calculated for *Frequent Users*. Search trails with at least two hops are used when measuring the staying power of a domain, and three hops (and two domains) when measuring the probability of returning to it. In addition to using our explicit (e.g., *Domain is Query*) and implicit (e.g., KL Preference) measures of preference to identify preferred domains, we also add a new implicit metric, *Frequent Visits*, to distinguish between preference and familiarity. While KL Preference and TF Preference are designed to capture a user's relative preference for a domain compared to others, *Frequent Visits* models the raw frequency of the visits to the domain. We consider *Infrequent Visits* to be visits where the user has visited the domain less than five times, and *Frequent Visits* more than five. Of domains frequently visited by the user, only 5.8% are not implicitly preferred (meaning they are popular across most users but not clicked more than average by the given user.). For infrequently visited domains, 22.0% are not implicitly preferred.

In general, we observe that if a user exhibits either an *Implicit* or *Explicit* preference for a domain, the user is more likely to stay within that domain. The user takes more steps within the domain, and is less likely to leave for a different domain while on the trail. If the user does leave the domain, that user is also more likely to return. The more frequently a user visits a domain, however, the more likely that user is to transition to other domains during the trail. This suggests that the set of frequently visited domains is probably different from the set of preferred domains. One explanation is that popular aggregator sites are visited frequently, and users then select new sites to visit from there.

The type of preference a user exhibits for a search result domain can indicate whether the user will return to the initial domain in a given *Search Result Trail*. When explicitly or implicitly targeting a domain, a user is more likely to return to the domain if the user leaves the domain in the search trail, but tends to return to the domain after a longer period with more hops. Users spend more hops in explicitly preferred domains than domains that are not explicitly preferred, yet slightly fewer hops in implicitly preferred domains than domains with no implicit preference. It may be the case that in explicitly preferred domains, users are browsing for new content, or it may be that they are seeking certain content such as a specific URL. In the former case, it may be advantageous to provide multiple search results from the preferred domain. In the latter case, a search engine interface could preemptively help the user arrive at that destination by displaying topical pages relative to the query, or the user's more common destinations from their preferred domain.

If a user has a preference for an initial domain in a search trail, the user is only slightly more likely to hop to the same second domain from that initial domain. If we consider domains that have been visited by the user before, 9.0% will lead to the same second domain hop from the same initial domain. For implicitly preferred initial domains, 9.8% will have the same second domain hop. For explicitly targeted domains, 10.9% will have the same second domain hop. On the other hand, 15.9% of frequently visited domains that are not implicitly preferred have the same second hop. This further supports the claim that familiarity in terms of frequency of visits is not the same as preference for a domain.

6. CONCLUSION

In this paper we studied individuals' domain preference in Web search. We described several methods to measure domain preference based on standard information retrieval techniques that account for the overall popularity of the domain and capture both positive and negative preferences. We found explicit and implicit domain preference to be both common and individualized among frequent search users. Implicit domain preference was relatively stable across topics and over time. The day to day preferences of a user did not tend to change, even during unusual events like holidays. We also showed that the implicit domain preferences people exhibited were stronger when searching than browsing.

It is apparent from our analysis that individuals exhibit personalized domain preferences, and that search engines and browsers have the ability to observe these explicit and implicit preferences by logging an individual's behavior. Many applications, such as search result ranking, advertising, news recommendation, and other Web-based information access and delivery services, could potentially benefit by using these preferences as input features into personalization algorithms. We hope that this work inspires the application of domain preference modeling in these areas.

7. REFERENCES

[1] Adamic, L.A. & N. Glance. 2005. The political blogosphere and the 2004 U.S. election: divided they blog. Workshop on Link discovery (LinkKDD '05).

[2] Adar, E., J. Teevan, & S. T. Dumais. 2008. Large scale analysis of Web revisitation patterns. *CHI '08*, 1197-1206.

[3] Bennett, P.N., F. Radlinski, R.W. White, & E. Yilmaz. 2011. Inferring and Using Location Metadata to Personalize Web Search. *Proc. SIGIR*, 135-144.

[4] Bharat, K. 2000. SearchPad: Explicit capture of search context to support Web search. *Proc. WWW*, 493-501.

[5] Bilenko, M. & R.W. White. 2008. Mining the search trails of surfing crowds: Identifying relevant websites from user activity. *Proc. WWW*, 51-60.

[6] Budzik, J. & K. Hammond. 1999. Watson: Anticipating & contextualizing information needs. *Proc. ASIST*, 727-740.

[7] Collins-Thompson, K. & P. N. Bennett 2010. Predicting Query Performance via Classification. In Advances in Information Retrieval (LNCS '10). Springer Berlin / Heidelberg, Volume 5993, 140-152.

[8] Dou, Z., R. Song, & J.R. Wen. 2007. A large-scale evaluation & analysis of personalized search strategies. *Proc. WWW*, 581-590.

[9] Elsweiler, D., D. E. Losada, J. C. Toucedo, & R. T. Fernandez. 2011. Seeding simulated queries with user-study data for personal search evaluation. *Proc. SIGIR,* 25-34.

[10] Glance, N.S., M. Hurst, K. Nigam, M. Siegler, R. Stockton, & T. Tomokiyo. 2005. Deriving marketing intelligence from online discussion. Proc KDD, 419-428.

[11] Glance, N.S., M. Hurst & T. Tomokiyo. 2004. BlogPulse: Automated trend discovery for weblogs. *WWW Workshop on the Weblogging Ecosystem: Aggression, Analysis and Dynamics '04*.

[12] Ieong, S., N. Mishra, E. Sadikov, & L. Zhang. 2012. Domain Bias in Web Search. *Proc. WSDM.*

[13] Jones, R. & F. Diaz. 2007. Temporal profiles of queries. ACM Trans. Inf. Syst. 25, 3, Article 14.

[14] Kim, J., K. Collins-Thompson, P.N. Bennett, & S.T. Dumais, 2012. Characterizing Web Content, User Interests, and Search Behavior by Reading Level and Topic. *Proc. WSDM.*

[15] Matthijs, N. & F. Radlinski. 2011. Personalizing Web search using long term browsing history. *Proc. WSDM*, 25-34.

[16] Morita, M. & Y. Shinoda. 1994. Information filtering based on user behavior analysis & best match text retrieval. *SIGIR '94*, 272-281.

[17] Obendorf, H., H. Weinreich, E. Herder, and M. Mayer. 2007. Web page revisitation revisited: Implications of a long-term click-stream study of browser usage. *CHI '07*, 597-606

[18] Pitkow, J., H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, & T. Breuel. 2002. Personalized search. *Commun. ACM*, 45, 9 (September 2002), 50-55.

[19] Piwowarski, B. & H. Zaragoza. 2007. Predictive user click models based on click-through history. *CIKM '07*, 175-182.

[20] Shen, X., B. Tan, & C.X. Zhai. 2005. Context-sensitive information retrieval using implicit feedback. *Proc. SIGIR*, 43-50.

[21] Shokouhi, M., R. White, P. Bennett, F. Radlinski. 2013. Fighting search engine amnesia: Reranking repeated results. *Proc. SIGIR*, 273-282.

[22] Silverstein, C., M. Henzinger, H. Marais, & M. Moricz. 1998. Analysis of a very large AltaVista query log. Technical Report 1998-014, Digital SRC.

[23] Singla, A., R.W. White, & J. Huang. 2010. Studying trail finding algorithms for enhanced Web search. *Proc. SIGIR*, 443-450.

[24] Speretta, M. & S. Gauch. 2005. Personalized search based on user search histories. *Proc. Web Intelligence*, 622-628.

[25] Sugiyama, K., K. Hatano, & M. Yoshikawa. 2004. Adaptive Web search based on user profile constructed without any effort from users. *Proc. WWW*, 675-684.

[26] Tauscher, L. and S. Greenberg. 1997. How people revisit Web pages: Empirical findings and implications for the design of history systems. *International Journal of Human Computer Studies*, 47(1), 97-138.

[27] Teevan, J., E. Adar, R. Jones, & M. A. S. Potts. 2007. Information re-retrieval: Repeat queries in Yahoo's logs. *Proc. SIGIR,* 151-158.

[28] Teevan, J., S.T. Dumais, & E. Horvitz. 2005. Personalizing search via automated analysis of interests & activities. *Proc. SIGIR,* 449-456.

[29] Teevan, J., D. J. Liebling, & G. R. Geetha. 2011. Understanding and predicting personal navigation. *Proc. WSDM*, 85-94.

[30] Tomokiyo, T. & M. Hurst. 2003. A language model approach to keyphrase extraction. ACL 2003 Workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18 (MWE '03), Vol. 18.

[31] Tyler, S.K. & J. Teevan. 2010. Large scale query log analysis of re-finding. *Proc. WSDM*, 191-200.

[32] Tyler, S.K., J. Wang, & Y. Zhang. 2010. Utilizing re-finding for personalized information retrieval. *Proc. CIKM*, 1469-1472.

[33] Weber, I. & C. Castillo. 2010. The demographics of Web search. *Proc. SIGIR*, 523-530.

[34] White, R.W. & J. Huang. 2010. Assessing the scenic route: Measuring the value of search trails in Web logs. *Proc. SIGIR*, 587-594.