Who Uses Web Search for What? And How?*

Ingmar Weber
Yahoo! Research Barcelona
Av. Diagonal 177
08018 Barcelona, Spain
ingmar@yahoo-inc.com

Alejandro Jaimes
Yahoo! Research Barcelona
Av. Diagonal 177
08018 Barcelona, Spain
ajaimes@yahoo-inc.com

ABSTRACT

We analyze a large query log of 2.3 million anonymous registered users from a web-scale U.S. search engine in order to jointly analyze their on-line behavior in terms of *who* they might be (demographics), *what* they search for (query topics), and *how* they search (session analysis). We examine basic demographics from registration information provided by the users, augmented with U.S. census data, analyze basic session statistics, classify queries into types (navigational, informational, transactional) based on click entropy, classify queries into topic categories, and cluster users based on the queries they issued. We then examine the resulting clusters in terms of demographics and search behavior. Our analysis of the data suggests that there are important differences in search behavior across different demographic groups in terms of the topics they search for, and how they search (e.g., white conservatives are those likely to have voted republican, mostly white males, who search for business, home, and gardening related topics; Baby Boomers tend to be primarily interested in Finance and a large fraction of their sessions consist of simple navigational queries related to online banking, etc.). Finally, we examine regional search differences, which seem to correlate with differences in local industries (e.g., gambling related queries are highest in Las Vegas and lowest in Salt Lake City; searches related to actors are about three times higher in L.A. than in any other region).

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human factors; H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Experimentation, Human Factors, Measurement

Keywords

query logs, demographics, session analysis, topic classification

1. INTRODUCTION

Web search is a multi-billion dollar industry and yet one of the biggest challenges is understanding the general characteristics of web searchers in terms of their demographic profiles, what they search for, and how they search. Unlike in traditional businesses, where customers come face to face with vendors at some point of the production and sales process, the web is unique because it can be used anonymously and search portals in particular act like gateways to a plethora of information and services across the globe. Analysis of user logs has therefore become an important research area because it forms the basis for many business decisions. The results can be used for SEO strategies, for improving user experience, for advertising, and for several other purposes. One area of particular interest has been the analysis of large-scale search query logs in order to understand user intent, topics being searched for, and quality of search results, among others.

In spite of all of the previous efforts to understand and model web users, there is still a lot to explore in terms of understanding large-scale web search behavior. In particular, in most studies demographics information has not been considered. The exception is data provided by services such as Alexa[1], Comscore[2], Quantcast[3] and others which provide basic demographic statistics for a range of websites. However, this data is mostly computed over individual web site visits, and does not include session analysis, nor search query topics. Marketing studies have also been done [10, 11] to gain insights into users' demographics, but these are also limited, and tools such as Yahoo! Web Analytics gather statistics only from the sites for which they are installed and do not provide the detailed analysis we carry out here.

In this paper, we analyze web search by joining behavioral observations and demographic features of users in aggregate. In particular, we aim to gain a deeper understanding of *who* is searching by analyzing users' demographics, of *what* they

*This research is partially supported by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037, "Social Media" (www.cenitsocialmedia.es). The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement nř215453 - WeKnowIt.

WSDM'11, February 9–12, 2011, Hong Kong, China.

[1] http:\\www.alexa.com
[2] http:\\www.comscore.com
[3] http:\\www.quantcast.com

are searching for by analyzing query topics, and of *how* they are searching, by analyzing session information. Our work differs from previous work in that we combine behavioral and demographic data to answer the questions above. In particular, we examine a large scale search engine's query log and analyze it along the dimensions of *who*, *what*, and *how*, and based on our analysis suggest that there are different types of searchers. The results of our classification has some similarities with traditional segmentations and creation of *personas* in Marketing. One key difference is that we perform our segmentation based on data analysis and automatic unsupervised (k-means) clustering.

The main contribution of our work can be summarized as follows: (1) analysis of a large-scale search query log combining behavioral and demographic information; (2) a classification of web searchers based on jointly on behavior and demographics. Our work provides important insights on search behavior and can lead to setting the basis for future work focusing on improved design decisions (e.g., should there be differences in the interface design for different types of searchers?), improved search results (e.g., can the search engines produce results that are better because of being aligned with demographic features [25]?), and further studies on information flows (e.g., how does information travel between different demographic groups and can this contribute to improving the information gap [26]?

2. RELATED WORK

Many researchers have tackled the problem of search query log analysis. Piwowarski et al. [20] classify web search activity models into three categories: (1) analysis models where the aim is to gain insights into typical user behavior; (2) models that try to predict the next user action, and (3) models that estimate the perceived relevance of a document. Our work falls mainly in the first category, but we briefly mention related work in all three areas.

Piwowarski et al. [20] built user activity models via Bayesian networks to predict the relevance of document search results for individual users. Manavoglu et al. [18] built probabilistic user models and create behavior model clusters in order to predict individual user actions, while White and Ducker [27] mined the query logs of 2,527 volunteers over a period of 5 months. Based on an analysis of the search query logs, they identified two classes of users: navigators (highly consistent search interaction) and explorers (highly variable search interaction). Demographic features were not used and query topics were examined only on a small sample of 385 queries.

Guo and Agichtein [8] extract a number of features, not just from the query log but also from more fine-grained user interactions such as mouse movements and query context in order to detect web searcher goals. Their work comprises a study with 10 subjects and an analysis of search logs from 440 users of a university's libraries. Guo et al. [9] also combine query, search, and interaction features, but to predict query performance. Wang et al. [24] use a Partially Observable Markov Model over query log data to model hidden variables and compare with eye tracking results. Kumar and Tomkins [17] study the behavior of Yahoo! toolbar users and propose a taxonomy of page views (news, portals, games, verticals, multimedia). Their analysis of search behavior classifies pages viewed after a search is issued, in terms of these categories, but the queries themselves are not classified and demographic data is not used.

Baeza et al. [1] presented a framework to identify users' interest using supervised and unsupervised learning. Queries are classified as informational, not informational, or ambiguous (after [3]), while Jansen et al. [13] automatically classified a much larger set of queries into informational, navigational, and transactional. The authors found that 80% of the queries in their dataset are informational in nature. Pu et al. [21] classified queries into categories in order to determine intention, while Beitzel et al. [2] analyzed hourly behavior differences in web search, and Bruce et al. [4] investigated the methods that people use in their workplace to re-access web information. They found that common refinding methods included the use of a search service and partial completion of a site's web address and acceptance of a suggested completion to this address (the auto-complete function).

Hu et al. [12], presented a technique to predict a search user's demographics, while Weber and Castillo [25] studied the relationship between individual queries and demographic variables. Our work differs in that we do not focus on prediction, and perform a larger scale and more in-depth query log analysis using demographic information than that of [25].

The market studies reported in [10] (based on enhanced focus groups[4] of 24 participants) and [11] (based on a survey of 425 participants) give general insights on web search behavior. In particular, from the focus groups four types of searchers were identified: (1) *Scan and clickers* (male, aged 23-28, average income $32,500, High-school-college education); (2) *2 step scanners* (male, aged 20-62, average income $47,900, College-University education), (3) *Deliberate searchers* (40% male, 60% female, aged 22-53, College-University education, average income $36,750), and (4) *1,2,3 Searchers* (20% male, 80% female, aged 22-41, average income $30,000, Collge level education). The survey also reports observations in gender variations in search behavior, as well as in types of queries (e.g., research vs. general type queries in product search) and other variables in relation to educational levels, income, age, and other demographic factors. In addition, generic consumer segmentations such as Mosaic's [6] exist. In particular, Mosaic segments UK consumers in terms of their socio-demographics, lifestyles, culture, and behaviour to classify consumers into 61 types aggregated into 11 groups. Such efforts are done by domain experts (30 in Mosaic's case).

One key difference between our work and previous marketing publications such as these is that we base our observations on analyzing a large-scale query log as opposed to doing surveys, working with focus groups, or building models based purely on domain knowledge. In practice, we think a combination of the four approaches brings the most benefit. To summarize, our work differs from previous research in the following aspects: (1) rather than focus only on keywords, we analyze sessions, classify queries, and use additional demographic data (e.g., election results); (2) we cluster users based on search topic categories and identify different types of behavior; and (3) we jointly examine three aspects of search behavior (who, what, how).

[4]A qualitative research methodology in which small groups of people (e.g., < 15 persons) are asked questions about a product, service, prototype, etc.

3. USER MODELING

We propose to analyze the behavior of web searchers along the following three orthogonal dimensions:

- Query topics ("what?"): What are the topics that the user issues queries on?

  In our study, we used the Y! Directory[5] classification as a definition of a "topic" but other topic definitions could be used or, instead of looking at topics, one could consider the fraction of queries having a local focus or containing product names.

- User demographics ("who?"): What is the demographic profile of the user?

  In our study, we used a mix of user-provided information (age and gender) and information derived from the user's ZIP code (expected income, expected educational level, expected political party affiliation). If available, other features such as their *actual* educational level, their marital status or even their body mass index[6] could be included.

- Session characteristics ("how?"): Does a user have many/few, short/long, and navigational/informational sessions?

  For this study, we limited ourselves to basic measures such as session length, number of queries per session or the fraction of sessions with suggested/guided queries. We also looked at the fraction of queries with a very high/low click entropy. If available, one could include features about the session's success or the user's frustration level [7].

Conceptually, this approach is very simple but, to the best of our knowledge, these dimensions have not been *simultaneously* studied. Note that the actual quantitative description in Section 5 might depend on the web search engine analyzed. However, we believe that the framework described here and the high-level description of the user groups generalize to other search engines and are good indications of general web search behavior. For the data used in our study (see Section 4 for details on how it was obtained), details on the "who?", "what?", and "how?" dimensions can be found in Tables 1, 2, and 3 respectively, including abbreviations for feature names that are used later.

4. DATA (PRE-)PROCESSING

The main data source for our study was a large sample of the web search query log of the Yahoo! search engine between 2008 and 2009. Queries were cast to lower case, white spaces at either end were removed and all other consecutive sequences of white spaces were replaced by a single "space" character. No other normalization, such as the removal of stop words or punctuation, was performed.

Our sample only included queries of registered U.S. Yahoo! users with an identifiable cookie who had provided demographic information (gender, birth year, country, ZIP code) upon registration. Furthermore, we only used the log for the U.S. Yahoo! site and only for users whose country in the self-provided information was the U.S.. We automatically removed users who claimed to have inexistent ZIP codes or who claimed to live in an area with 0 population (as of the U.S. 2000 Census[7]). Finally, our sample only included "active" users, that is, those who had issued at least 100 queries over the sample period. To filter out additional bots, on top of a proprietary bot filtering algorithm, we also removed users with more than 100,000 queries and users who, on average, had clicked on organic results for fewer than 1/100 of their queries or more than 100 times per query. Our final sample thus contained queries for 2.3 million registered users, for which no personal information was used. All analysis was anonymous and performed in aggregate.

[5] http://dir.yahoo.com/

[6] `http://en.wikipedia.org/wiki/Body_mass_index`

[7] http://factfinder.census.gov/

Feature (abbreviation)	mean	stderr	10%-ile	90%-ile
birthyear (b-yr)	1968	15	1948	1986
gender (g)	49.0	50.0	-	-
per-capita income (pc-i)	22.9k	9.6k	14.3k	33.5k
below poverty (bp)	10.9	7.9	3.2	21.4
av. household size (hhs)	2.6	0.4	2.3	3.1
mean travel time (mtt)	25.7	6.0	18.4	33.4
non-english (ne)	17.1	17.4	3.2	41.7
ba degree (ba)	25.7	15.1	9.8	48.6
black (bl)	10.2	16.7	0.4	28.7
white (wh)	77.5	21.0	46.5	97.0
asian (as)	3.9	7.0	0.2	9.6
hispanic (his)	11.7	16.8	0.9	33.8
obama (ob)	52.7	13.9	35.0	70.0
mccain (mc)	46.1	13.9	29.0	64.0

Table 1: Demographic values for our data set of 2.3 million active web search users. Most fields were obtained by joining the user-provided ZIP code with U.S. census information as described in Section 4.1.

4.1 "Who?" Data

In order to understand better *who* these users are and how that relates to *what* they search for, we used the ZIP code to obtain estimates of demographic information such as per-capita income, level of education, and ethnicity from the U.S. 2000 Census (see [25]). We also included 2008 U.S. presidential election results in our analysis. For this we mapped county names to ZIP codes[8] and then obtained the election results for each county[9]. Table 1 gives a complete list of the demographic features we extracted for each user. Note that for all of the ZIP-derived demographic information we labeled users according to their expected (fractional) values. For example, a user from the ZIP code 94087 would be labeled as 61.3% white, as 30.3% asian and as 1.6% black.[10]

4.2 "What?" Data

The second dimension of our study was "what are people searching for?" and by this we refer to the topic distribution of the queries issued by a user.

First, the top 10 Yahoo! web search results were obtained for a given query. Then, for each of the ten resulting web pages we obtained the unique classification by a proprietary

[8] `http://www.getzips.com/county.htm`

[9] `http://elections.nytimes.com/2008/results/president/votes.html`

[10] Our terminology for the race and the other demographic features is the same as in the official U.S. 2000 Census. See `http://factfinder.census.gov/home/en/epss/glossary_r.html#race` for details.

Category	abbr.	Five most frequent queries in category	%-age queries vol./dist.	macro av. mean/std.
Arts - Humanities	a/hum	barnes and noble, twilight, amazon books, ancestry.com, borders	1.6/2.8	1.8/3.4
Bus. & Econ. - Employment & Work	b/emp	careerbuilder, indeed, monster.com, monster, indeed.com	1.1/1.3	1.2/3.6
Bus. & Econ. - Finance & Investment	b/fin	bank of america, chase, paypal, wells fargo, chase.com	4.7/4.2	4.5/8.1
Bus. & Econ. - General	b/gen	ebay, walmart, amazon, home depot, ebay.com	4.3/2.5	4.4/6.3
Bus. & Econ. - Shopping & Services	b/sho	craigs list, craigslist.com, qvc, realtor.com, macys	5.3/5.7	5.4/7.6
Computers & Internet - General	c/gen	facebook login, yahoo mail, facebook.com, gmail, yahoo.com	4.4/2.3	4.2/8.4
Education - General	ed/gen	fafsa, classmates, university of phoenix, classmates.com, sallie mae	2.0/3.3	2.1/5.0
Entert. - Consumer Electronics	en/con	best buy, verizon wireless, verizon, att, at&t	2.8/2.6	3.1/4.9
Entert. - Movies & Film	en/mov	netflix, imdb, robert pattinson, blockbuster, movies	2.5/2.7	2.7/4.6
Entert. - Music	en/mus	myspace, my space, amazon.com, taylor swift, project playlist	5.2/5.2	4.9/8.1
Entert. - Televion Shows	en/tv	tmz, hulu, perez hilton, cartoon network, comcast	5.1/4.5	5.0/7.2
Health - Diseases & Conditions	h/dis	webmd, web md, swine flu symptoms, swine flu, mayo clinic	1.2/2.0	1.6/3.3
News & Media - General	n/gen	facebook, craigslist, youtube, google, yahoo	14.3/4.4	11.4/14.3
Recreation - Automotive	rec/aut	ebay motors, autotrader, autozone, cars.com, kelley blue book	2.5/3.7	2.9/5.8
Recreation - Games	rec/gam	pogo, pogo.com, adicting games, comcast.net, runescape	4.1/3.9	3.9/9.0
Recreation - Sports	rec/spo	espn, nfl, epsn.com, nfl.com, nascar	3.8/5.0	3.7/7.0
Recreation - Travel	rec/tra	southwest airlines, expedia, travelocity, orbitz, american airlines	2.8/3.1	3.2/5.7
Reference - General	rec/gen	white pages, yellow pages, wikipedia, dictionary, ask.com	1.3/0.8	1.6/3.2
Science - Biology	sci/bio	club penguin, petsmart, petfinder, petco, clubpenguin	1.3/2.3	1.5/3.4
Science - Geography	sci/geo	mapquest, google maps, mapquest driving directions, map quest, maps	1.0/0.4	1.2/2.7
Society & Culture - Food & Drink	soc/foo	food network, pizza hut, papa johns, recipes, domino's pizza	2.7/4.4	3.2/4.8
Society & Culture - Relationships	soc/rel	tagged, match.om, plentyoffish, wal mart, adam4adam	2.1/1.1	1.6/6.3
Society & Culture - Sexuality	soc/sex	youporn, redtube, pornhub, xnxx, free porn	7.1/5.5	6.4/13.9

Table 2: Details for search query topics, each corresponding to at least 1% of the query volume. These 23 topics cover 83% of the total volume. The table shows the five most frequently issued queries (always navigational), and (i) the micro-averaged percentages (in terms of categorized query volume and categorized distinct queries) and (ii) the macro-averaged percentages (where topic distributions are first computed for each user and then averaged across the 2.3 million users). See Section 4.2 for details.

classifier[11], truncated to include only 71 classes found in the the first two levels of the Y! Directory (`http://dir.yahoo.com/`). Finally, we used a majority voting scheme over the ten results, where the result at rank k has a vote with weight $11 - k$. In case of ties the algorithm checks on whether the two strongest classes have a common parent class. If that is the case, the parent class is used (e.g., Health/General is used for Health/Fitness and Health/Nutrition), otherwise the algorithm indicates that the class is unknown.

Queries are usually short and might have multiple meanings, thus our approach has two advantages: (1) classifying web pages is easier than classifying queries as the former have more content and structure, and (2) since classification is done on the results, we leverage the fact that results are optimized by the search engine for a particular query. Thus, the classification takes advantage of the structure of the web as well as any other information used by the search engine ranking algorithm (this could include among others, measures of quality of the results, relationship between queries and clicks to a particular result aggregated over millions of users, etc.).

We classified only queries issued at least 30 times, and this yielded 1.0 million distinct queries. For each user the query topic distribution was then re-normalized such that the sum over the 71 topics was 100%. Since infrequent queries were not classified and tail queries might not be distributed evenly across the topics, we also ran experiments *without* this re-normalization. However, the qualitative results obtained, both for the user segmentation (Section 5) and for the inter-dependence between the features (Section 6), were very similar and we only report results for the re-normalized setting. Table 2 gives an overview of some topics that accounted for at least 1% of the classified query volume.

Although the classifier worked well for the majority of queries, there are some details that need to be pointed out. First, the topic News&Media/General is dominated by the queries *facebook*, *craigslist*, *youtube*, *google* and *yahoo*, which can be regarded as being related to changing information or media, but which are probably not what one would intuitively expect in this category. Second, most frequent queries in the topic Computer&Internet/General are related to popular web sites such as `mail.yahoo.com` or `mail.google.com`, which might be counter-intuitive. Finally, the vast amount of query volume in Science/Geography refers to online map applications such as `mapquest.com`.

[11]These were the top 10 results as of Mid-June 2010. Changes in the query results between the time the queries were actually issued and the time of classification were *not* considered.

4.3 "How?" Data

The third and last dimension of our study refers to the "how?" of web search by which we mostly refer to session features such as session length and the number of clicks per session. As a definition of a session we did *not* apply any intent-based segmentation [14] but used a simple timeout interval of 30 minutes, which was also used in [8]. The session length is the time between the first query within the interval and either the last query within the interval or, if the last query led to click results, the result click corresponding to a query issued within the interval boundaries. For each session we counted the number of queries issued and we also recorded if any of the queries was the result of a clicked/selected query suggestion, including the "explore concepts" (suggested query) option. We refer to such queries as "guided queries". Table 3 gives a complete list of the "how?" features as well as some (macro) statistics about their distributions across users. By macro we mean that, for example, when reporting averages we first obtained one number for each user, which in the case of certain features such as average session length already required averaging for each user, and then we considered the averages across all users for the corresponding feature.

We also classified queries into navigational, informational, and transactional. In order to distinguish between navigational queries such as "facebook" and informational queries such as "health risks of smoking" we broke down queries according to their click entropy. If query q results in a click on document d and $p(d|q)$ is the fraction of times that d was clicked out of all cases when *some* document was clicked in response to q, then the click entropy $H(D|q)$ for query

q is defined as $H(D|q) = \sum_{d,q} -p(d|q) \log_2 p(d|q)$. Note that according to this definition any query which resulted in a total of n_c clicks could have a maximum click entropy of $\log_2 n_c$, meaning that infrequent queries would trivially have a low click entropy. In order to avoid this inherent bias, we chose to compute the click entropy only for queries which were issued at least 20 times. We then labeled queries with $H(D|q) \leq 1.0$ as "focused" queries and queries with $H(D|q) \geq 3.0$ as "diverse". In both cases, there were around 370k distinct such queries. Table 4 lists frequent exemplary queries, indicating that "focused" queries tend to be navigational whereas "diverse" queries tend to be informational or, in the case of adult content, transactional. In the computation of the click entropy, as well as in our study in general, we only used clicks on "organic" (i.e., non-sponsored), algorithmic search results and did *not* include clicks on advertising as the statistics on such clicks are business-sensitive.

Feature	mean	std.	10%-ile	90%-ile
# sessions (s-c)	103	78	45	187
time/session (sl)	405	180	221	611
queries/session (qps)	2.4	.7	1.7	3.2
frac. single click sess. (fscs)	.41	.13	.25	.57
frac. guided sessions (fgs)	.16	.11	.04	.30
organ. clicks/session (cps)	2.0	.8	1.2	2.9
% focused (foc)	19.3	17.1	3.4	43.4
% diverse (div)	9.5	7.7	2.4	17.8

Table 3: "How do people interact with web search engines?": This table shows the distribution, macro-averaged across users, for basic session features. Focused/diverse refers to low and high click entropy queries respectively. Details on how this entropy is computed and on the definition of a session can be found in Section 4.3.

4.4 Averages and Statistical Significance

All statistics we report are macro per-user statistics. As stated earlier, when reporting averages we first obtained one number for each user. In the case of certain features such as average session length this required averaging for each user, and then we considered the averages across users in the corresponding cluster or feature decile. Only the column referring to the percentage of queries in Table 2 refers to micro-averages, averaged over all query instances for which a topic classification was obtained.

Some words on statistical significance: as the set of users considered in this study comprises 2.3 million users, even the slightest differences between two non-random subsets of users tends to be significant at levels of 1% or lower. To see this consider the following example. Suppose you have a set of 2 million normally distributed values with a sample average of 0.0 and a sample standard error of 1.0. Now, you split the 2 million values into two halves according to some meta information such that in one half the average is $\mu_1 = -.003$, whereas in the other equally-sized half it is $\mu_2 = +.003$. Intuitively, one might expect that the mean in both halves should still be indistinguishable, or rather not differ statistically significantly from the background mean of 0.0 as the difference of .003 is small compared to the standard error of 1.0. However, this is not the case! When averaging $1,000,000$ random variables distributed $N(0,1)$ the final result is a random variable with a distribution of $N(0, 1/\sqrt{1,000,000})$ as the standard error drops as $1/\sqrt{n_k}$ where n_k is the size of the cluster under consideration. Now for this tiny standard error the two means mentioned above differ both from each other and from 0.0 at a level of statistical signficance below 1%.

In short, *all* the differences reported in this paper are statistically significant at levels below 1%, but we chose to present only differences which we considered "interesting", where interesting means that the relative differences in means should be noticeable.

Focused		Diverse	
Query	Cl. ent.	Query	Cl. ent.
amazon	0.22	american idol	4.46
best buy	0.41	baby names	3.03
craigslist	0.19	dancing with the stars	5.66
ebay	0.77	games	3.23
espn	0.86	jobs	3.38
facebook	0.80	michael jackson	7.16
mapquest	0.71	myspace layouts	3.91
myspace	0.21	robert pattinson	6.89
target	0.21	twilight	4.92
yahoo mail	0.54		

Table 4: A list of some frequent "focused" queries with click entropy ≤ 1.0 (left) and some frequent "diverse" queries with click entropy ≥ 3.0 (right). Focused queries tend to be navigational while diverse queries tend to be informational or, in the case of adult queries, transactional. Click entropy was only computed for queries which were issued at least 20 times. Details in Section 4.3.

5. WEB SEARCH USER SEGMENTATION

Using the general approach from Section 3, in this section we present a user segmentation of active web search users, similar in spirit to typical market segmentations such as those created by Nielsen or Mosaic [6].

In order to automatically obtain such a segmentation we clustered users using their representation in the "what?" dimension, and then investigated what kind of segmentation in the other two dimensions were induced. We chose this approach, rather than (co-)clustering all three dimensions simultaneously, as the topical interests of a user are most relevant for disambiguating queries or to serve relevant content, including advertising. However, as we will see, all three dimensions are closely intertwined and users with different topical interests also differ in terms of the "who?" and the "how?". In Section 6 we explicitly investigated the interplay between the "what?" and the "who?", between the "what?" and the "how?", and between the "who?" and the "how?".

The detailed segmentation presented in Table 5 and discussed next was computed using a k-means clustering algorithm [15, 16] for the topic distributions for each user.[12]

[12]We used the implementation available at `http://www.cs.umd.edu/~mount/Projects/KMeans/` and ran k-means clustering with 300 iterations, using all four variants available in the software ("Lloyd's", "Swap", "EZ_Hybrid" and "Hybrid"). The lowest mean-squared distance was obtained for

As far as the number of clusters is concerned we tried different values of k ranging from 8 to 20. The qualitative differences between the centroids differed little but, when a smaller value of k was chosen then clusters with ID 4 and 7 (see Table 5 – both with a large fraction of transactional Society&Culture/Sexuality queries) and clusters with ID 3 and 10 (both with a large fraction of navigational News&Media/General queries) were merged into a single cluster. In all cases, independent of the k, each centroid could however be reasonably clearly assigned to one of the following three groups with the same high-level description of the "what?", "who?" and "how?".

5.1 The Informational Users

Historically, evaluation setups such as the "Cranfield Paradigm" [23] had their origin in the library sciences. The search engine's role was to serve as an improved and automatic indexing scheme, pointing the user towards relevant articles. The "typical" user in this setting is envisioned to issue exclusively, or at least predominantly, informational queries. In our segmentation we call a user an "informational user" if he differs from the average user in that he is more likely to use a web search engine to find information on a wide range of topics. Put simply, these users use web search engines as a web *re*search engines. As far as the "how" is concerned this means that (i) they are more likely to issue non-navigational queries, (ii) they are less likely to have single-click sessions, (iii) they are more likely to make use of the suggested query alternatives. Along the "what" axis, these users are diverse and do "research" on a wide range of topics, with little interest in adult content. In terms of their demographic profile, these users are more likely to be well-educated and have an above-average income. See clusters with IDs 5, 13, 12, 3 and 0 (sorted in descending order of size) in Table 5 for more details.

5.2 The Navigational Users

As the quality of web search engines has improved, bookmarks have lost some of their importance and a large fraction of queries are re-finding queries [22]. Table 2 shows that in all of the topics the most frequent queries are navigational where the user might just as well type the query in the browser's address bar to re-find a URL from his browsing history. We call a user "navigational" if he differs from the average user in that he is more likely to use a web search engine to navigate to URLs that he already knows exist. Put simply, these users use web search engines as a replacement for web page bookmarking. As far as the "how" is concerned this means that (i) they are more likely to issue navigational queries, (ii) they are more likely to click only on a single result within a session, (iii) they are less likely to make use of the (unnecessary) suggested query alternatives. The "what" axis for these users is dominated by the topics of popular websites such as News & Media (Facebook, Craiglist or Youtube - cluster IDs 10 and 3), Computers & Internet (Yahoo Mail, Gmail - cluster ID 6), Entertainment/Music (myspace, project playlist - cluster ID 14), Recreation/Sports (cluster ID 15). In terms of demographics, the averages are close to the background averages, with certain variations depending on the topical cluster under consideration.

the "Hybrid" variant and we only report quantitative results for this setting.

5.3 The Transactional Users

Unlike classical paper documents, web pages are often interactive and allow transactions, such as the purchase of items or the download of video clips. We call a user "navigational" if he differs from the average user in that he is more likely to use a web search engine to take him to *some* URL where he can perform the desired transaction. They differ from the informational users in that a single result will "do the job" and there is little benefit in learning more about a subject. They differ from the navigational users in that the result URL is generally not known in advance and there are several alternatives to choose from. In terms of "what?" predominant topics are shopping, adult content and gaming sites. In terms of "how?", queries tend to have diverse clicks, i.e. a high click entropy, and the interaction with the search engine tends to be short compared to informational queries. The "who?" depends heavily on the kind of transaction with Society&Culture/Sexuality (clusters with ID 4 and 7) attracting predominately men whereas Business&Economy/Shopping (cluster with ID 8) is dominated by women. Transactional users with an interest in Recreation/Games (cluster with ID 11) are more likely to come from neighborhoods with below average income and educational level.

5.4 Close-up on Selected Clusters

Whereas the description in the previous three sub-sections was on a rather high level, here we zoom-in on a few selected clusters and describe them in a manner similar to [6] and [5]. Cluster names are deliberately oversimplified to emphasize the relative differences between the clusters.

- Baby Boomers[13] (Cluster ID 1): Users in this cluster tend to be older than the typical web searcher with an average age of 50 years. Their predominant topic of interest is finance and a large fraction of their sessions consist of simple navigational queries related to online banking.
- Adult Content Seekers (Cluster ID 4): A large fraction of these mostly male users' queries are of a transactional kind involving adult content. They are slightly older than the average user and are often "satisfied" with a single click in a session. Also see Cluster ID 7.
- Liberal Females (Cluster ID 8): The typical user in this group is female and more likely to have voted Democrat in the 2008 elections. The biggest single query topic is shopping and sessions are comparatively long, hinting at possible browsing and comparison behavior.
- White Conservatives (Cluster ID 12): Users in this group are more likely to be white and live in areas that voted Republican in the 2008 elections. Mostly male, they often search for automotive related topics, business pages and, compared to the average, relatively often for home and garden information.
- Challenged Youth (Cluster ID 14): These users are comparatively young with an average age of 34. They tend to live in low-income neighborhoods with a low level of education. Their searches are centered around music and their sessions are often of a navigational kind. Also see Cluster ID 13.

[13]The U.S. Census Bureau defined Baby Boomers as those born between 1946 and 1964.

5.5 Comments

Note that we do *not* claim that this segmentation is strict in the sense that an "informational user" never issues navigational queries, just as the query *michael jackson* can have both an informational or a transactional intent. We do however claim that the mix of the three categories are not the same for all users and that for the different user types observed the "how?", the "what?" and the "who?" are related.

Note that, in absolute numbers, the differences for ZIP-derived features, such as income or educational level, are small but statistically significant (see Section 4.4). However, we believe that one should *not* interpret this to mean that these factors have little influence on the web search behavior of a user. Rather the differences between, say, people with or without university degree are partly washed away as the educational level can only be observed at the aggregate ZIP-code level and *not* at the level of each individual. Furthermore, the census data used is 10 years old and many neighborhoods will have changed since then. In short, we expect the discovered trends to be much more pronounced given up-to-date per-person estimates and the trends observed in the present study can give valuable pointers on where it might be valuable to obtain such data, e.g. through questionnaire-based user studies.

6. FEATURE INTERDEPENDENCE

The user segmentation discussed in Section 5 was derived from a clustering using *only* the differences in the topic distribution since the "what?" is most relevant for advertising and for serving relevant content. Still, even this segmentation revealed correlated differences along the other two dimensions. In this section, we explore these interpendences in more detail. Recall that, as explained in Section 4.4, due to the large sample size all differences listed below are statistically significant, even if they appear small.

Note that we do *not* list correlations within each dimension though. For instance, regions with a low per-capita income also tend to be regions with a low educational level, users interested in Health/Reproductive Health are also more likely to be interested in Society&Culture/Families, and users with a large fraction of single click sessions are less likely to click on suggested queries in any given session.

6.1 Interplay between "What?" and "Who?"

Looking at the query topic distribution, there is clear evidence that "what women want" is not the same as "what men want." Exemplary differences can be found for the topics of Sexuality (f: 2.8%, m: 10.1%) and Sports (f: 2.4%, m: 4.9%), which are more popular among male users, and for Reproductive Health (f: 0.3%, m: 0.1%), Arts/Crafts (f: 0.6%, m: 0.2%) and Families (f: 0.9%, m: 0.4%), all three of which are more popular among female users.

Similarly, as illustrated in Figure 1, certain topics depend heavily on the user's birth year[14]. Some topics overpronounced among older users include Health/Diseases & Conditions (h/dis), Gambling (rec/gamb), and Travel (rec/tra). People in their late 20s are the group most interested in Health/Fitness (h/fit) and Reproductive Health (h/rep), whereas young people show the biggest interest among age groups in Games (rec/gam) and Education/General (ed/gen).

Apart from these "primary" demographic features we also observed a notable dependence on other ZIP-derived features. Based on the per-capita income estimates all of the following topics are most popular among the lowest income decile of users. Entertainment/Music (en/mus), Entertainment/ Comics & Animation (en/comi), and Government/Military. The percentage of people of Asian decent is greater for Computers & Internet/Programming & Development, with Asian neighborhoods appearing more active than others. Finally, we also observed that the following topics were over-expressed compared to the average user in predominantly white neighborhoods: News & Media/Weather, Recreation/Outdoors, Society & Culture/Home & Garden, and Science/Agriculture.

To summarize, gender and birthyear are the strongest demographic features when it comes to differentiating between topic interests and the differences are stereotypical. But other features also reveal interesting differences. For instance, users in "very" white neighborhoods are about three times as likely to issue queries related to Home & Garden or Agriculture than users in "very" non-white neighborhoods.

6.2 Interplay between "What?" and "How?"

Given that a user has many short "navigational sessions" with a single click, his query topic distribution differs from a user who has fewer such one-click sessions. Similarly, users who rarely or never click on suggested queries have different topical interests than users who make use of such guidance help. Depending on the fraction of single click sessions, the percentage of a user's queries related to the following topics changes noticeably. All of News & Media/General, Society& Culture/Relationships and Computers & Internet/General are most popular among users with many single click sessions but, interestingly, they are still more popular among users with the lowest fraction of such sessions when compared to users with an intermediate fraction. There is a trend in the opposite direction, with the *lowest* interest by people with a large fraction of single click sessions, for the topics Health/ Diseases & Conditions, Health/Medicine, Social Science/General and Arts/Humanities. These behavioral differences indicate that topics such as health or arts tend to correspond to non-trivial, information-seeking sessions.

A similar trend can be observed when one investigates which topics users, who are more/less likely to click on query suggestions, are interested in. Concretely, the fraction of guided sessions influences the most percentage of topical queries in the following categories: Society & Culture/Sexuality, Society & Culture/Relationships, Health/ Diseases & Conditions, Health/Nutrition, Science/ Biology and Social Science/General.

To summarize, users with a large fraction of multi-click sessions and users who are more inclined to click on query suggestions tend to be more interested in "informational topics" such as health, science or arts, and less interested in transactional adult content topics or in navigational News & Media/General topics, which are dominated by sites such as Facebook and Craigslist.

6.3 Interplay between "Who?" and "How?"

Who you are not only influences what you are searching for online but also *how* you search for it. For instance, a

[14]Absolute values are business sensitive, so they are omitted. However, we highlight differences that are statistically significant as explained in Section 4.4, even if those differences are not obvious in Figure 1 or not quantitatively reported.

ID	size	"What?"	"Who?"	"How?"	Type
0	50,224	edu 25.7^{+}, h/med $.9^{+}$, n/gen 6.7^{-}, soc/sex 1.6^{-}	g 33^{-}, b-yr 1971^{+}, bl 13.2^{+}, wh 74.3^{-}	div 6.9^{-}, se-c 93^{-}, qps 2.5^{+}, cps 2.1^{+}	I
1	74,303	b/fin 38.1^{+}, en/tv 2.1^{-}, soc/sex 1.9^{-}, n/gen 6.7^{-}	b-yr 1960^{-}	div 5.9^{-}, fscs $.45^{+}$, fgs $.08^{-}$, foc 23.3^{+}	N
2	124,436	en/tv 26.7^{+}, en/mov 7.4^{+}, n/gen 6.2^{-}, b/gen 2.3^{-}	b-yr 1971^{+}, g 44^{-}	div 14.8^{+}, foc 13.3^{-}, fgs $.18^{+}$, qps 2.5^{+}	I
3	269,444	n/gen 26.1^{+}, soc/sex 2.5^{-}	g 43^{-}	foc 26.1^{+}, div 7.3^{-}, fscs $.43^{+}$	N
4	73,586	soc/sex 64^{+}, b/sho 1.6^{-}, n/gen 4.0^{-}, en/con 1.2^{-}	g 85^{+}, b-yr 1966^{-}	div 18.0^{+}, fgs $.09^{-}$, fscs $.45^{+}$, qps 2.3^{-}	T
5	672,335	r/tra 5.6^{+}, soc/foo 5.5^{+}, h/dis 2.6^{+}, soc/sex 1.7^{-}	g 39^{-}, ba 27.6^{+}, p-ci $23.9k^{+}$	foc 12.1^{-}, fgs $.18^{+}$, fscs $.39^{-}$	I
6	55,464	c/gen 45.1^{+}, b/sho 3.0^{-}, en/tv 2.2^{-}, soc/sex 2.0^{-}	b-yr 1962^{-}, g 44^{-}	foc 31.6^{+}, div 5.5^{-}, fgs $.11^{-}$, fscs $.47^{+}$	N
7	162,010	soc/sex 29.1^{+}, r/tra 2.1^{-}, b/sho 3.4^{-}, soc/foo 1.9^{-}	g 79^{+}, b-yr 1970^{+}	div 13.5^{+}, foc 17.2^{-}	T
8	81,935	b/sho 34.4^{+}, r/toy 4.6^{+}, soc/hol 1.6^{+}, soc/sex 1.3^{-}	g 28^{-}, ob 54.1^{+}, mc 44.6^{-}	sl 446^{+}, se-c 93^{-}, foc 21.6^{+}, div 8.7^{-}	T
9	24,428	soc/rel 51.2^{+}, b/gen 1.7^{-}, soc/foo $.9^{-}$, h/dis $.4^{-}$	g 60^{+}, p-ci $21.5k^{-}$, bl 13.3^{+}, b-yr 1966^{-}	foc 42.8^{+}, fgs $.08^{-}$, s-c 130^{+}, qps 2.1^{-}	N
10	114,598	n/g 57.3^{+}, b/gen 2.2^{-}, h/d $.6^{-}$, soc/foo 1.1^{-}	b-yr 1966^{-}	foc 45.6^{+}, fgs $.09^{-}$, div 4.7^{-}, fscs $.48^{+}$	N
11	48,331	r/gam 49.6^{+}, b/gen 1.8^{-}, b/sho 1.8^{-}, r/tra 1.0^{-}	ba 23.0^{-}, p-ci $21.5k^{-}$	fscs $.48^{+}$, fgs $.14^{-}$, cps 1.78^{-}, sl 353^{-}	T
12	168,924	r/aut 14.0^{+}, b/gen 15.4^{+}, en/con 5.2^{+}, soc/hom $.5+$	g 62^{+}, b-yr 1964^{-}, wh 80.3^{+}, mc 47.8^{+}	se-c 94.5^{-}, fscs $.39^{-}$, foc 17.4^{-}, div 8.6^{-}	I
13	190,283	r/gam 12.8^{+}, en/comi 2.0^{+}, en/mov 5.2^{+}, en/con 4.2^{+}	b-yr 1974^{+}, ba 24.1^{-}, p-ci $21.9k^{-}$	foc 11.9^{-}, fgs $.20^{+}$, div 11.6^{+}, qps 2.5^{+}	I
14	86,121	en/mus 35.3^{+}, b/fin 2.0^{-}, soc/foo 1.6^{-}, soc/sex 3.5^{-}	b-yr 1976^{+}, b-p 12.8^{+}, p-ci $20.6k^{-}$, ba 21.9^{-}	foc 27.8^{+}, fscs $.43^{+}$, div 8.5^{-}	N
15	63,701	r/spo 33.5^{+}, n/gen 7.8^{-}, b/sho 3.2^{-}, c/gen 2.2^{-}	g 78^{+}, b-p 10.0^{-}, ba 28.3^{+}, p-ci $24.3k^{+}$	fgs $.14^{-}$, div 8.1^{-}	N

Table 5: Details of the centroids for k-means clustering when users are clustered using *only* their topic distributions. But as the "what?" is not independent of the "who?" and the "how" (see Section 6), statistically significant differences compared to global per-user macro-averages were also found in the other dimensions. The $^{+}$ and $^{-}$ indicate that a certain feature is over- or underexpressed compared to the corresponding cluster-independent macro-average. The type refers to (I)nformational, (T)ransactional or (N)avigational users as explained in Section 5. See Table 2 for topic abbreviations.

larger percentage of people holding a bachelor degree in a ZIP code area means that the fraction of focused queries with a low click entropy tends to be lower ([0-9.8]%: 20.4, [21.8-25.9]%: 19.3, [48.6-100]%: 17.3) but the same holds for diverse queries with a high click entropy ([0-9.8]%: 10.3, [21.8-25.9]%: 9.5, [48.6-100]%: 8.3). People in educated neighborhoods also have a smaller fraction of guided sessions ([0-9.8]%: 0.17 [21.8-25.9]%: 0.16, [48.6-100]%: 0.15) and their average sessions are shorter ([0-9.8]%: 418s, [21.8-25.9]%: 406s, [48.6-100]%: 394s).

In short, people with higher educational levels tend to have shorter sessions, tend to be less inclined to click on query suggestions and, it seems, are more likely to issue infrequent tail queries. This last observation is based on the fact that for the decile with the highest fraction of bachelor degrees the fraction of *both* focused and diverse queries, which are queries with a minimum of 20 occurrences, are lower than the background distributions. We plan to investigate this phenomena in more detail in the future, as well as details related to the use of advanced query operators.

Examining users from different birth year decile ranges, one can also note the following differences in the fraction of diverse queries, with younger people issuing a larger fraction of such queries (['00-'48]: 9.0, ['70-'73]: 9.1 ['87-'06]: 11.6). Correspondingly, the fraction of focused queries also changes with the birth year (['00-'48]: 19.6, ['70-'73]: 20.0 ['87-'06]: 16.4). The observation that older users seem to issue a larger fraction of focused, navigational queries was also made in [25] where it was noted that they have a larger fraction of queries containing the token *www*. Details of how the birth year affects the session behavior are given in Table 6.

Birth year	# sess	fscs	fgs	qps	div	foc
1900-1948	103	.41	.14	2.3	19.6	9.0
1970-1973	106	.42	.16	2.4	20.0	9.1
1987-2006	93	.40	.21	2.6	16.4	11.6

Table 6: A breakdown of selected session characteristics for three birthyear deciles. Interestingly, young users have a larger fraction of "guided sessions", involving at least one click on a suggested query.

7. REGIONAL DIFFERENCES

So far we only looked at the "who?" in terms of demographics such as age, gender or (expected) income. In this section we report certain findings on *regional* differences. This was partly inspired by the work in [19] where regional food preferences are detected from the query log of a recipe web site. In order to split users according to regions, we broke up our user population by the first two digits of their self-reported ZIP code. Such prefixes tend to agree with either metropolitan areas, where the prefix 10* roughly corresponds to New York City, or in some cases even small states, where the prefix 58* covers all of North Dakota.[15]

Table 7 shows differences in the topical query distributions for a selection of ten different two-digit ZIP prefixes. The table shows that, for example, the fraction of searches related to actors is about 3 times higher in the L.A. area, which includes Hollywood, than in any other of the regions considered. Similarly, the fraction of queries related to gambling is highest in Las Vegas and lowest in Salt Lake City.

Note that most of the differences in Table 7 seem to correlate with differences in the dominant industry in each region. As information about the economic characteristics and, in particular, the distribution of the work force across industries is part of the census information, we plan to include this information in future work. This would allow, for example, a distinction between rural areas where agriculture is important and high tech regions, where IT is dominant.

8. CONCLUSIONS

In this paper we presented an analysis of a large-scale search query log of anonymous registered users from a major web search engine. We jointly studied on-line behavior in terms of *who* these users might be (demographics), *what* they searched for (query topics), and *how* they searched (session analysis). We used information provided by users during registration (age, sex, zip code) along with data obtained from the US census, and classified queries into topic categories, analyzed basic session statistics, and clustered users based on the topics of the queries they issued. Our anal-

[15]See `http://en.wikipedia.org/wiki/ZIP_code_prefixes` for details.

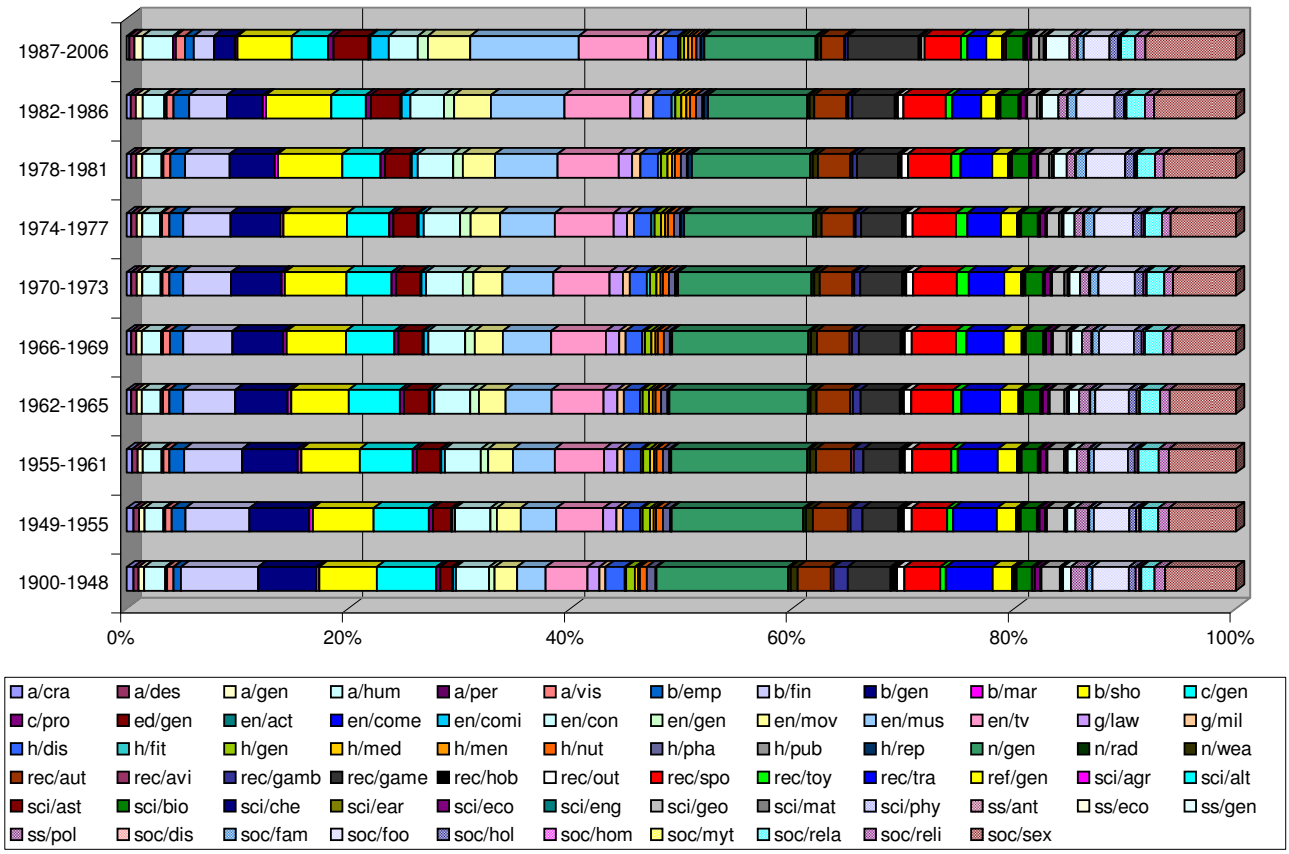

Figure 1: Query topic distribution broken down for birth year deciles. For older users the fraction of finance (b/fin) and travel (rec/trav) related queries is higher. For younger users the fraction of music (en/mus) and gaming (rec/game) related queries is higher.

ysis showed that by examining the resulting clusters it is possible to identify distinct patterns of behavior along different demographic features. We were also able to classify query topics into informational, navigational, and transactional categories based on click entropy, and upon examining the clusters obtained we were able to suggest different types of searchers (e.g., Baby Boomers, White Conservatives, etc.) in terms of the topics they search for and their behavior. Overall, the findings are "stereotypical" with men more interested in adult content than women, with health topics attracting more of a research behavior (rather than navigational behavior), and with people with higher educational levels issuing fewer navigational queries (which are arguably more efficiently typed in the browser's address bar directly), etc..

Although our study comprises analysis of data for only one search engine, the results do give us important insights on web search in terms of analyzing all three dimensions jointly (who, what, how). Future work includes more fine grained analysis in terms of categories and search strategy, as well as integration with additional information (e.g., on local industries, unemployment, etc.), and a closer look at long-tail queries.

9. REFERENCES

[1] R. Baeza-Yates, L. Calderón-Benavides, and C. González-Caro. The intention behind web queries. In *String Processing and Information Retrieval (SPIRE)*, pages 98–109, 2006.

[2] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. Hourly analysis of a very large topically categorized web query log. In *Conference on Research and development in information retrieval (SIGIR)*, pages 321–328, 2004.

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

[4] H. Bruce, W. Jones, and S. Dumais. Keeping and re-finding information on the web: What do people do and what do they need? *Proceedings of the American*

ZIP prefix	Region	# users	co/prog	en/actors	h/medi	n/radio	r/gambl	r/trav	sci/agricult
10***	New York City, New York	15.2k	.39	.01	*.43*	.25	.80	3.83	.08
25***	Charleson, West Virginia	5,954	.31	.00	.40	.18	.49	2.34	.17
58***	North Dakota	3,182	.48	.00	.31	.22	.53	2.74	.34
69***	North Plate, Nebraska	895	.37	.01	.25	*.35*	.64	2.97	*.59*
82***	Wyoming	3,039	.32	.01	.27	.25	.51	3.56	.23
84***	Salt Lake City, Utah	11.4k	.46	.01	.31	.23	.31	3.29	.12
86***	Flagstaff, Arizona	4,228	.36	.01	.34	.25	.85	3.63	.10
89***	Las Vegas, Nevada	20.0k	.46	.01	.26	.23	*1.09*	3.92	.08
90***	L.A., California	54.8k	.49	*.03*	.39	.29	.67	3.52	.06
94***	San Fransisco, California	56.8k	*.62*	.01	.38	.21	.55	*4.36*	.07

Table 7: Regional differences in query topic distributions for a selection of ten regions and seven topics.

Society for Information Science and Technology (JASIST), 41(1):129–137, 2004.

[5] ESRI. Tapestry segmentation, 2010. `http://www.esri.com/library/fliers/pdfs/tapestry_segmentation.pdf`.

[6] experian. Mosaic united kingdom - the consumer classification for the uk, 2009. `http://www.ccr.co.uk/pdf/MOSAICGuide.pdf`.

[7] H. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *Conference on Research and development in information retrieval (SIGIR)*, pages 34–41, 2010.

[8] Q. Guo and E. Agichtein. Ready to buy or just browsing?: detecting web searcher goals from interaction data. In *Conference on Research and development in information retrieval (SIGIR)*, pages 130–137, 2010.

[9] Q. Guo, R. W. White, S. T. Dumais, J. Wang, and B. Anderson. Predicting query performance using query, result, and user interaction features. In *Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO)*, 2010.

[10] G. Hotchkiss. Inside the mind of the searcher. Report, Enquiro, March 2004.

[11] G. Hotchkiss. Search engine usage in north america. Report, Enquiro, April 2004.

[12] J. Hu, H. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *Conference on World Wide Web (WWW)*, pages 151–160, 2007.

[13] B. Jansen, D. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of Web queries. *Information Processing & Management (IPM)*, 44(3):1251–1266, 2008.

[14] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Conference on Information and knowledge management (CIKM)*, pages 699–708, 2008.

[15] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu. A local search approximation algorithm for k-means clustering. In *Symposium on Computational geometry (SOCG)*, pages 10–18, 2002.

[16] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(7):881–892, 2002.

[17] R. Kumar and A. Tomkins. A characterization of online browsing behavior. In *Conference on World wide web (WWW)*, pages 561–570, 2010.

[18] E. Manavoglu, D. Pavlov, and C. L. Giles. Probabilistic user behavior models. In *International Conference on Data Mining (ICDM)*, page 203, 2003.

[19] J. Miller. Butterballs or cheese balls, an online barometer, 2009. `http://www.nytimes.com/2009/11/26/dining/26search.html`.

[20] B. Piwowarski, G. Dupret, and R. Jones. Mining user web search activity with layered bayesian networks or how to capture a click in its context. In *Conference on Web Search and Data Mining (WSDM)*, pages 162–171, 2009.

[21] H. Pu, S. Chuang, and C. Yang. Subject categorization of query terms for exploring Web users' search interests. *Journal of the American Society for Information Science and Technology (JASIST)*, 53(8):617–630, 2002.

[22] J. Teevan, E. Adar, R. Jones, and M. A. S. Potts. Information re-retrieval: repeat queries in yahoo's logs. In *Conference on Research and development in information retrieval (SIGIR)*, pages 151–158, 2007.

[23] E. Voorhees. The philosophy of information retrieval evaluation. In *Workshop of the Cross-Language Evaluation Forum (CLEF)*, pages 143–170, 2002.

[24] K. Wang, N. Gloy, and X. Li. Inferring search behaviors using partially observable markov (pom) model. In *Conference on Web search and data mining (WSDM)*, pages 211–220, New York, NY, USA, 2010. ACM.

[25] I. Weber and C. Castillo. The demographics of web search. In *Conference on Research and development in information retrieval (SIGIR)*, pages x–x, 2010.

[26] I. Weber and A. Jaimes. Demographic information flows. In *Conference on Information and knowledge management (CIKM)*, pages 1521–1524, 2010.

[27] R. W. White and S. M. Drucker. Investigating behavioral variability in web search. In *Conference on World Wide Web (WWW)*, pages 21–30, 2007.