The Demographics of Web Search

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

Carlos Castillo
Yahoo! Research Barcelona
Diagonal 177, 08018 Barcelona, Spain
chato@yahoo-inc.com

ABSTRACT

How does the web search behavior of "rich" and "poor" people differ? Do men and women tend to click on different results for the same query? What are some queries almost exclusively issued by African Americans? These are some of the questions we address in this study.

Our research combines three data sources: the query log of a major US-based web search engine, profile information provided by 28 million of its users (birth year, gender and ZIP code), and US-census information including detailed demographic information aggregated at the level of ZIP code. Through this combination we can annotate each query with, e.g. the average per-capita income in the ZIP code it originated from. Though conceptually simple, this combination immediately creates a powerful user modeling tool.

The main contributions of this work are the following. First, we provide a demographic description of a large sample of search engine users in the US and show that it agrees well with the distribution of the US population. Second, we describe how different segments of the population differ in their search behavior, e.g. with respect to the queries they formulate or the URLs they click. Third, we explore applications of our methodology to improve web search relevance and to provide better query suggestions.

These results enable a wide range of applications including improving web search and advertising where, for instance, targeted advertisements for "family vacations" could be adapted to the (expected) income.

Categories and Subject Descriptors

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

General Terms

Human Factors, Measurement

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

Keywords

Web search, Demographic factors

1. INTRODUCTION

What kind of web results would you personally want to see for the query "`wagner`"? Well, if you are a typical female US web user you probably have pages about the composer Richard Wagner in mind.[1] However, if you are a male US web user you are more likely to be referring to a company called Wagner which produces paint sprayers.[2] Similarly, the term most likely to complete the beginning "`hal`" is in general "`lindsey`"[3], whereas for people living in areas with an above average education level the most likely completion is "`higdon`"[4] These two examples illustrate that demographic factors have a measurable influence on search behavior.

Even though this user modeling can be an interesting end by itself, the ultimate goal for the search engine is to provide a better service to users. This can take the form of more relevant search results, more helpful query suggestions, more interesting news items on a portal page or, last not least, improved targeting of advertising. Of course, the potential for improvements in these areas is not new and there exists a rich set of literature on the use of *personalization* [6, 26].

Our approach of grouping users by demographic features, such as age or income, is different from personalization where models are learned for individual users. Given enough data about individual users, personalization can indeed be very powerful. However, our approach is more applicable in practice because (i) it requires less training data, as we use a small number of different user classes, (ii) it has less potential for breaching the user's privacy, as we do not aggregate information on a per-user level, and (iii) it has a simpler interpretation, as we use real-life features with clear semantics instead of latent variables.

As far as sponsored search is concerned, especially this last issue is of relevance as advertisers could, e.g., choose to have their ads displayed only for users in a particular expected income range. This would go well beyond the current "demographic site selection", which uses only age and gender

[1]For women the most clicked URL was `http://en.wikipedia.org/wiki/Richard_Wagner`.
[2]For men the most clicked URL was `http://www.wagnerspraytech.com/`.
[3]Hal Lindsey is an American evangelist and Christian writer.
[4]Hal Higdon is an American writer and runner.

and works only for banner ads, but is not widely used for sponsored search.[5]

The main contributions of our study are:

- we demonstrate how public information for ZIP codes can be used to annotate both web queries and URLs with demographic features;
- we show that the user population of a large, commercial search engine is representative of the whole US population;
- we uncover differences in search behavior across demographic segments, e.g. in the form of "representative" queries; and
- we show that using demographic information has a potential to improve state-of-the-art web search results, especially for difficult queries, and that it leads to improvements in query suggestions.

The rest of this paper is organized as follows. The next section outlines previous work related to ours. Section 3 describes how we obtained and processed the data. Section 4 discusses our methodology. Basic characteristics of the demographics of web searches are shown in Section 5. The impact of using demographic information on three different application scenarios, (i) web search, (ii) automatic labeling of URLs, and (iii) query suggestions, is quantified in Section 6. As we see our work as a first, exploratory study, we discuss possible extensions in Section 7. The final section presents concluding remarks.

2. RELATED WORK

Inferring demographics from behavior. It is well established that the writing style of texts can be used to infer characteristics of their authors, such as gender or age [1, 2]. Koppel et al. [23] apply this method to blog pages.

Search logs kept by search engines have been used to infer demographic characteristics of their users. Hu et al. [12] represent users by the words in the pages they click on, and by the output of a topic classifier on those pages. They are able to determine gender with 80% accuracy and age (discretized in 5 ad-hoc age groups) with 50% accuracy. Jones et al. [15] represent users by the queries written by them. They achieve 84% accuracy on gender and 79% accuracy on age (within ±10 years of error). In the current paper we use an approach that is related to both: we represent users by the queries they write and by the identity (but not the contents) of the documents they click upon.

Gender and internet usage. Jackson et al. [13] interviewed 600 undergraduate students and found that in general females used e-mail more and web less than males ("*Women communicating and men searching*" is the sub-title of their article). Gavin et al. [14] interviewed over 600 students and found that males used the internet in general more than females, but found no evidence of a different pattern of communication vs search among genders. Specifically in the case of web search, Lorigo et al. [17] use eye-tracking to uncover differences in search strategies across different types of tasks and among genders.

[5] `http://adwords.google.com/support/aw/bin/answer.py?hlrm=en&answer=33743`

Income, race, and the digital divide. When demographic dimensions such as income or age are related with internet usage, an unequal access is observed, which is known as the Digital Divide [22]. The digital divide, for instance, establishes a gap between countries with a high gross domestic product and countries with a low gross domestic product in terms of broadband access. The digital divide also can be observed within a country by itself. For instance, in the US it has been observed that more income and more education are correlated with a higher probability of having a computer at home, but that African Americans[6] are much less likely to have a computer at home across all levels of income and education [10].

The need for demographic-aware search. Morgan and Trauth [20] advocate for taking into account individual differences in web search. Ford, Miller and Moss [7, 9, 8] have studied this topic extensively, emphasizing the differences among the strategies people with different cognitive styles use to understand information and acquire knowledge. For a survey of works on search behavior, see [11].

Groupization to improve search. Teevan et al. [27, 21] investigated the benefit of using "groupization" rather than personalization to improve web search ranking. In their studies, done in a work environment with a total of roughly 200 participants, they showed that pooling search history information from members in the same "group", ranging from job function to gender, can lead to gains in normalized Discounted Cumulative Gain (DCG), in particular for group-related queries. Apart from the setting (work environment with access to desktop information vs. anonymized use of query logs), and the scale ($<$ 1,000 users vs. $>$ 1,000,000 users), one of the main differences to our work is that we are also interested in descriptive results (see Section 5), as well as applications apart from web search (see Section 6).

Conditional entropy of search logs. The question of how much could we improve web search by knowing the identity, or rather the IP address, of a web user, was studied in [19]. Specifically, they measure the conditional entropy $H(URL|Q, IP)$ and compare it to $H(URL|Q)$. Though these are interesting and intuitive measures to look at, there are certain kinds of methodological problems in practice. For example, even when tuples (URL, Q, R) for a random variable R with $H(R) = H(I)$ were considered, where R is *independent* of both Q and URL, then for a fixed pair (Q, R) the number of matching tuples (URL, Q, R) would be small. This means that the distribution of URL within a "bucket" of a given value of R will naturally look homogeneous, as the bucket is far to small to represent an adequate sample of the random variable URL. Therefore, conditioning on *any* additional variable X, be it IP or R, will *always* lead to a drop of $H(URL|Q, X)$ compared to $H(URL|Q)$, even if X is independent of both URL and Q. This problem could potentially be avoided by looking at the *cross entropy*, i.e. by learning the distribution of the triples (URL, Q, IP) on one set of training data, and then using these probabilities to estimate $H(URL|Q, IP)$ on a different set of test data. However, this approach is also problematic as a single URL in the test set, which was unseen in the training set,

[6] We use the term "African American" as a short-hand for the term "black or African-American" as used in the official US 2000 census. See `http://factfinder.census.gov/home/en/epss/glossary_r.html#race`.

would yield an *infinite* entropy estimate. Ignoring any such unseen URLs will, on the other hand, lead to a biased *underestimate* of the actual entropy. These issues could be addressed by comparing the the empirical conditional entropy $H(URL|Q, IP)$ to $H(URL|Q, R)$ where $H(IP) = H(R)$ for an *independent* random variable R. But even then the results would remain difficult to interpret, as it is not clear, for instance, how much a reduction in conditional entropy of 5% would affect the web search results as the entropy might be reduced without changing the *ranking* of URLs. Hence, we preferred to look at a more tangible measure, the precision at one. Our methodology is described in detail in Section 4.

3. DATA ACQUISITION AND PROCESSING

For this study, we used the following three sources of information. First, a subset of the query log data for US search traffic of the Yahoo! web search engine. Second, profile information (birth year, gender and ZIP code) provided by registered users. Third, publicly accessible demographic information for US ZIP codes, obtained in the 2000 census[7], and joined with the other data sources on the ZIP code (explicitly provided by users).

3.1 Query Log Preprocessing

The collected data went through the following cleaning and preprocessing steps. First, only web searches issued between October 1, 2008 and September 30, 2009 were considered. Second, we used a subset of the queries in this time window, sampled uniformly at random. Third, only web searches by users logged into our services were used, and to protect users' privacy, all user identifiers were hashed using a non-invertible function. Furthermore, during the whole process, nothing was ever aggregated on a per-user level, which is one of the advantages of our proposed approach.

Fourth, only web searches (i) originating from the US-version of the search engine, and (ii) pertaining to users with a valid ZIP code were used.[8] Fifth, queries without clicks on URLs were discarded. We denote the pairs of query and clicked result URL by (query, URL) and when multiple URLs where clicked for the same query, multiple such pairs were generated. Sixth, queries were cast to lower case but no stemming was applied and all special characters (such as apostrophes) were kept. Seventh, immediate, repeated duplicates of (query, URL) pairs by a single user were conflated to a single instance. Note that we still kept repeated (query, URL) pairs for a single user as long as there were other pairs in between.

The final preprocessing step depended on the final analysis intended and we considered (input, target) pairs of different types. Concretely, in our first set of experiments we analyzed factors predicting the clicked URL (target) for a given query (input). For our second set of experiment, we looked at differences in how different queries (target) are used to describe the same URL (input). For our third set of experiments, we investigated potential improvements for "suggest as you type" interfaces and we only used queries which started with a sequence of non-white characters of minimum length two (input) followed by another such sequence (target). In the vast majority of cases these sequences were actual terms, but tokens including special characters were also kept. In all three settings, we only used (input, target) pairs where the input had a *support* of at least two users, i.e. there where at least two matching (query, URL) pairs (after the initial screening) originating from different users.

The remaining (input, target) pairs were then labeled with demographic information derived either directly from the user's profile (birth year and gender) or derived from using demographic information pertaining to ZIP codes. Details of this annotation process are described in the following section. The sizes of our final data sets are given in Table 1.

Input	→	Target	Pairs	Distinct inputs	Distinct users
query	→	URL	479 M	22.7 M	28.3 M
URL	→	query	588 M	41.7 M	29.7 M
1st term	→	2nd term	509 M	1.6 M	25.3 M

Table 1: Basic statistics for the preprocessed query log used for our study. In all cases, each input was used by at least two distinct users.

3.2 Demographic Feature Extraction

Apart from the the birth year and the gender, which were directly provided by the users, we also used the provided ZIP code to annotate each (input, target) pair with additional information. Concretely, we obtained the average values for the following features for each ZIP code:

- per capita income (in 1999 US dollars) [P-c income k$],
- bachelor's degree or higher, for population 25 years and over [BA degree %],
- individuals below poverty level [below poverty %],
- race: white [white %], black or African American [African American %], Asian [Asian %],
- speaks a language other than English at home, for population 5 years and over [non-English %].

In all cases, the full name is as in the official census information and, in brackets, is a shorthand used by us. For our current study, we decided to limit ourselves to this subset of features, but in future studies we might also include other available data such as "mean travel time to work in minutes" or "average family size". Note that the demographic features are *not* independent and, e.g., areas where a large percentage of the population holds a BA degree tend to have a higher per capita income as well.

The labels applied to each (input, target) pair were discretized. For all demographic features we used quintiles: the percentile intervals $[0\%, 20\%]$, $(20\%, 40\%]$, ..., $(80\%, 100]$. E.g., a ZIP code with no more than 12.8% of its population 25 years and over holding a bachelor's degree would be placed in the lowest quintile for the corresponding feature and, similarly, users born after 1982 would be in the youngest quintile as described in Table 2. Only for (i) the gender, where only two buckets were used, and (ii) the ZIP code itself, where we used only the two leading digits giving a total of 99 buckets, we did not use quintiles.[9] As we were *not* interested in merely geographical differences, we only

[7]This data is freely available at `http://factfinder.census.gov/`.

[8]Some of the user-provided zip codes were non existent, e.g. 12345, or they corresponded to an area without any registered inhabitants, e.g. 01244.

[9]The ZIP prefix 09 refers solely to locations outside the US, but all other prefixes from 00 to 99 were present. See `http://en.wikipedia.org/wiki/ZIP_code_prefixes`.

used the shortened ZIP code to filter out localized queries from the lists in Table 3 but did not use them for anything else. All percentiles were derived from (query, URL) pairs and the corresponding buckets were used in all experiments, even if the percentiles for other (input, target) combinations would have differed slightly. Percentiles were computed on a per query instance basis and *not* on a per-user basis.

The reason for this discretization is that we explicitly wanted to *de*-personalize and *de*-localize our analysis. E.g., using the full birth year could potentially isolate a very small user group and the same holds for using the *exact* average per-capita income, which would correspond to using the full zip code. Given even larger amount of query log data, one could investigate the differences in web search behavior between users born in, say, 1978 and in 1977, or the differences for the zip codes 95967 and 95969. However, one of the strengths of our approach is exactly the fact that we work with more *abstract* features, such as "very young" (youngest 20%) and "very old" (oldest 20%) or "very rich" or "very poor", thereby allowing a more intuitive interpretation.

	Query-log data					US
Feature	20%	40%	60%	80%	avg.	avg.
P-c income k$	16.0	18.9	22.4	27.7	22.7	21.6
Bel. poverty %	4.5	7.2	10.9	16.5	11.1	12.4
BA degree %	12.8	18.1	25.6	37.6	25.5	24.4
White %	61.9	78.8	88.1	94.4	76.9	75.1
Afr. Amer. %	0.9	2.4	5.7	15.5	4.0	12.3
Asian %	0.4	1.1	2.3	5.1	4.0	3.6
Non-English %	4.5	7.9	14.0	27.3	17.3	17.9
Year of birth	1956	1966	1974	1982	1968.7	1974
Gender	49.7% female		50.3% male		49.1% vs 50.9%	

Table 2: Aggregated per-query demographics from our query-log data, compared to the US average from census data (last column).

Table 2 summarizes the per-query demographics in our dataset, and compares them with averages in the US population, shown in the last column.[10] Although the averages agree for most features, ZIP codes with a high percentage of African American population appear to be underrepresented in our data set. This is consistent with findings on the "Racial Divide" [10]. For birth year, 1974 in the table for US population, only the *median* and not the (lower) mean birth year was available for the US population. As infants are no web users the average for our data set is expected to be below the US average.

3.3 Data Quality

Though it is certain that a fraction of users provided false profile information, sometimes deliberately, and that some of this will have escaped the test for the validity of a ZIP code, our results (Sections 5 and 6) show that the remaining signal is still significant. In settings with poorer data quality or without any profile information, approaches to automatically learn a user's profile from her search history seem promising [15].

Instead of using the user-provided zip code, we could have derived the ZIP code by mapping the user's IP address to a geographical location.[11] We did, however, not experiment with this approach as we expect that (i) the accuracy of the mapping is not high enough (as a "poor" ZIP code can be geographically neighboring a "rich" ZIP code) and that (ii) the mistakes could have a systematic bias (as internet providers could tend to have their routers in neighborhoods of a particular demographic profile). Still, if sufficiently accurate, such mappings could be used to extend the applicability of our work significantly by enabling a "demographic profiling" for requests to arbitrary web servers. Information derived from IP addresses could, when the IP-derived location differs significantly from the actual location, also be used to remove false or outdated ZIP code information. For the present study, we opted *not* to do this and rather to work with the "raw" data instead.

Note that we do not claim that web users are always representative of their area. Trivially, any individual user could always deviate in an arbitrary way from the "typical" resident of her neighborhood. But even the aggregated averages might not be representative. E.g. in a poor neighborhood the majority of its web users could be made up by people who are, by the local standards, financially better off than their ZIP code would imply. We are only claiming that despite such drawbacks our derived "labels" are still useful for (i) eliciting typical differences (see Section 5) and for (ii) predicting user behavior (see Section 6).

4. METHODOLOGY

The main objective of our experiments is to measure how much demographic information helps to improve the ranking of *targets*, e.g. clicked URLs, for a given *input*, e.g. a given query. More formally, let X, Y and D be random variables corresponding to the input, target and demographic information respectively. Similarly, let x, y and d be actual instances of values of these random variables. In this formulation we want to know how often the *mode* of the distribution changed, i.e. how often $\text{argmax}_y P(y|x,d) \neq \text{argmax}_y P(y|x)$. For both conditional distributions we required the two top ranked values to correspond to *distinct* probabilities, not counting cases of ties as improvements in ranking.

In other words, we looked at the improvement of the "precision at one" (P@1), assuming that (i) the search engine ranks according to the empirical click probability, which is almost always the case[12], and that (ii) a click is an indication of relevance so that P@1 is identical to the click-through-rate for the value with the highest (empirical) click probability.

For our evaluation, we only considered cases with sufficient demographic coverage to possibly allow a re-ranking. Concretely, we only considered inputs x with a support, in the probabilistic sense, of at least 100 users for some some combination (x, d) as well as at least 400 users for other values of d. See Table 6 for experimental results for this setup. The alternative possibility of using conditional entropy [3] was discussed in Section 2. Note that we do make use of the conditional entropy for *descriptive* purposes, in particular in Section 5, but we do not use it for performance evaluation.

[10]Details about the US averages and the definitions of the features can be found at `http://factfinder.census.gov/`

[11]See e.g. `http://www.hostip.info/` for an open community-based project to map an IP address to geographical locations.

[12]Of course, there is also a feedback loop and the highest ranked results will get clicked more often, independent of their relevance.

5. BASIC DEMOGRAPHIC DIFFERENCES

We begin by presenting partly anecdotal evidence describing how different demographic groups differ in their web search behavior. Table 3 lists the four most "discriminating" queries for different demographic groups, such as very young people, or people living in predominantly "white" neighborhoods. Queries are ranked by the average feature value, where the average is on a per-query basis; adult queries are not included.

If this were the only ranking criterion, then the list would be dominated by localized queries from single neighborhoods with highly skewed demographic distributions. E.g. the "richest" query with a support of at least 16 users would be "`paws for life`" (145k$, 43 occurrences, 28 users) and the "most educated" "`www.diamondbacks.com`" (73.3% BA degree, 333 occurrences, 22 users). The first refers to an animal charity in Maryland and the second to a baseball team in Arizona. As such examples fail to illustrate the full potential of our approach, for Table 3 we imposed the additional filter that each query must satisfy $H(\text{short ZIP}|\text{query}) \geq 4.0$, where short ZIP refers to the first two digits of a five digit ZIP code. Note that $H(\text{short ZIP}) = 6.23$ so that a query with $H(\text{short ZIP}|\text{query}) = 4.0$ covers about $2^{4.0}/2^{6.23} = 21\%$ of the US. Note that the entropy constraint also leads to the requirement of at least 16 distinct users for each query. The lists are generated such that "given the query, the demographics is determined", and *not* the other way around.

Many of the queries in Table 3 agree with stereotypes. For example, queries predominantly issued by young users tend to be related to chat rooms, music and social networking sites. Queries which are issued exclusively by male users in our sample are related to sports, or computer hard- and software. Queries from areas where a language other than English is often spoken at home, turn out to be written in Spanish. The "`s2s magazine`" query, found in the list for the feature "African American", refers to a magazine "covering the world of black entertainment"[13]. The query "`tvb series`", found in the list for the feature "Asian", refers to a series from a TV station based in Hong Kong.

In some cases the entries are less expected. E.g. "`www.unitnet.com`" is often issued as a web query from ZIP codes where a substantial amount of individuals lives below the poverty line, 26.4% compared to 12.4% US average. This site is the portal page for the "directors" –salespeople– of Mary Kay, a brand of skin care and color cosmetics which has been criticized by its opponents as a "product-based pyramid scheme" [5]. Also unexpected is the fact that many queries indicative for older people relate to annual shareholder meetings.

Though not our main focus, we also report on observations relating to differences in the actual *search process*. Table 4 shows that there is a slight but detectable trend for people with a university degree, or at least living in areas where this is more common, to type longer queries, to click more often on a single result, and to click on "deeper" URLs, meaning that they contain more "/"'s. *Query length* refers to the number of sequences of non-white-spaces. *URL depth* refers to the number of parts of a URL separated by a "/" and not including the protocol. E.g. "`sigir's homepage`" has query length 2 and `http://www.sigir.org/events/events-upcoming.html` has URL depth 3. As for each of the three groups in Table 4 the numbers are computed over roughly 95.8M = 0.20 × 479M (query,URL) pairs, these differences, though small, are statistically significant at a confidence level well below 0.001, using a t-test for equality of means.

[13] `http://www.dmoz.org/Society/Ethnicity/African/African-American/News_and_Media/Magazines/`

Feature	Query	Value
Per-capita income k$	chris jordan	81k
	electric candle warmer	78k
	www.popsugar.com	75k
	ns4w.org	65k
below poverty line %	www.unitnet.com	26.4
	slaker	25.8
	kipasa	24.9
	www.tokbox.com	24.4
BA degree %	spencer stuart executive search	55.5
	insight venture partners	54.2
	federal circuit	53.2
	four seasons jackson hole	52.8
White %	pulloff.com	97.1
	central boiler wood furnace	96.2
	firewood processors	96.1
	midwest super cub	95.5
African Americ. %	trey songz bio	63.8
	def jam records address	58.4
	s2s magazine	58.1
	madinaonline	56.0
Asian %	sina	25.1
	big bang lyrics	24.3
	tvb series	24.2
	jay chou lyrics	23.5
Non-english lang. %	mis novelas favoritas	60.5
	sinonimos	59.2
	juegos para baby shower	54.5
	dichos mexicanos	54.3
birth year, "old"	www.johnshopkinshealthalerts.com	1931
	www.envisionreports.com/vz	1935
	yahoo free bridge games	1935
	bnymellon.mobular.net/bnymellon/frp	1935
birth year, "young"	free teen chatrooms	1991
	wet seal	1991
	tottaly layouts	1990
	photofiltre brushes	1990
gender, female %	scrapbook myspace layouts	100
	eyeshadow for brown eyes	100
	twilight movie screensavers	100
	plus size jewelry	100
gender, male %	2009 nfl team rankings	100
	football big board	100
	resharper	100
	radeon x600	100

Table 3: Some highly-discriminating queries for our demographic features. Note that these queries, though discriminative in terms of $P(D|Q)$, are not necessarily typical for a demographic group in the sense of $P(Q|D)$. For illustrative purposes we show queries with a nation-wide focus by requiring a minimum entropy of 4.0 over the first two ZIP digits.

% BA degree	query length	click entropy	URL depth
Lowest 20%	2.25	1.74	1.92
Middle 20%	2.28	1.67	1.95
Top 20%	2.32	1.60	1.99

Table 4: People that are more likely to have a university degree (based on where they live), (i) type longer queries, (ii) click more often on a single result URL, and (iii) click on "deeper" URLs.

As can already be seen in Table 3, older people tend to be more likely to use URLs as web queries. Out of all 15.7M (query, URL) pairs where the query started with "`www.`" 29% were due to the "oldest" quintile, indicating an over-representation of this group for this behavior. Note that such queries are generally more appropriately placed directly in the URL bar of a browser. This can be seen as an indication that older people are on average less experienced with respect to web search.

It is also informative to look at examples of queries q where a demographic group d has an unusually high or low conditional click entropy $H(U|q,d)$ [3]. A high click entropy can be due to a number of reasons. It can be that the presented web results for that query are poor and people have to try many pages, but it can also be seen an expression of high interest on a potentially multi-faceted topic. For example, the query "`scrapbooking`" has a general click entropy of 5.4. But for the youngest quintile, the click entropy *increases* to 6.8, despite the fact that smaller buckets usually lead to smaller entropies. The same phenomenon holds for the oldest quintile and the query "`civil war`".

Gender G	$H(Q\|G)$	$H(U\|Q,G)$	$H(U\|G)$	$H(Q\|U,G)$
Male	19.12	1.87	19.85	2.61
Female	19.04	1.83	19.75	2.77

Table 5: Basic differences in how women and men search the web. The first set of columns is computed for the (query,URL) data set and the last two columns for the (URL,query) data set. See Table 1. Both for queries and URLs the uncertainty/diversity for men is about $2^{.09} = 6.4\%$ higher than for women, even though they show less variability for queries used to describe clicked URLs.

6. APPLICATIONS

The previous section already demonstrates that there is indeed a difference in search and click behavior between different demographic groups. In this section, we investigate if these differences can be exploited to improve web search (Section 6.1), automatic labeling of URLs (Section 6.2) and query suggestions based on query completions (Section 6.3).

6.1 Web Search

As we mentioned in the introduction, for the query "`wagner`" women predominantly click on the Wikipedia page about the composer and men on the page of a producer of spray brushes. Similarly, most people searching for "`esl`" click on the homepage of "ESL Federal Credit Union", but for people in areas with many households where a non-English language is spoken at home, the preferred result is a page with background information about the "English as a Secondary Language" exam. Such examples illustrate that knowing a particular demographic feature such as gender can potentially help to improve web search results. Here we go beyond anecdotal evidence and quantify the attainable improvement gains.

Applied to all queries with sufficient support (see below), the gains in P@1 are small. This is mostly due to the fact that the baseline system, a large commercial web search engine, already has a highly tuned ranking for common queries. Therefore, we also looked at more difficult subsets where the uncertainty of the clicked URL for a given query was large, i.e. where the empirical conditional entropy $H(U|Q)$ was above a certain threshold. For these cases, the demographic information is more useful to explain the diversity in the clicked URL.

Application	Instances	Precision @ 1 Baseline	Ours	Gain
§6.1 Search	207 M	.703	.713	1.4%
§6.1 Search $H(U\|Q) \geq 1.0$	123 M	.557	.574	3.0%
§6.1 Search $H(U\|Q) \geq 2.0$	61 M	.381	.408	7.1%
§6.2 URL labeling	246 M	.461	.483	4.8%
§6.3 Query completion	459 M	.250	.276	10.4%

Table 6: Improvements in P@1 by exploiting demographic information, for the different applications described in Sections 6.1-6.3 . We selected instances (input, target) where the input has a support of at least 100 users for *some* demographic feature value d, as well as at least another 400 users for other values of the same feature. P@1 was computed for those instances. The baseline system ranks targets according to $P(y|x)$. Our system ranks them by $P(y|x,d)$. The last column shows the relative gain.

Only cases with sufficient support to reliably use demographic data were used for Table 6. In detail, we required at least 100 users for a particular demographic feature value, such as a quintile for the age, as well as at least 400 users for other values of the same demographic feature. Even though only less than half of the (query,URL) pairs in our original data set (see Table 1) satisfy this criterion, we believe that in practice this fraction can be increased by more aggressive pre-processing of query terms, e.g. (i) by removing special characters (e.g. conflating "`men's health`" and "`mens health`"), (ii) by stemming (e.g. reducing "`cheap flights`" and "`cheap flight`" to singular form) and (iii) by grouping semantically related keys (e.g. combining "`ny times`" and "`new york times`" into one query).

6.2 Automatic URL Labeling

In the previous section, we looked at (query,URL) pairs and tried to predict the URL for a given query. Here, we look at the dual setting, i.e. given a clicked URL, can we predict the query it was clicked for?

Looking at frequently used queries leading to a particular URL essentially helps to give a concise description of a web page, very similar in spirit to using tags to label a page [18].

These descriptions might not be the same across different demographic groups. For instance, a marijuana-related page is found by most people using the slang "weed", whereas people in the oldest quintile use queries with the term "marijuana" to locate the *same* page. Apart from being used as labels, the *distribution* of the queries gives a hint at the function or status of a particular page. E.g., a popular site such as Facebook has a very low entropy in terms of the queries used to find it. As it is almost exclusively "found" by navigational queries, it could be called a "navigational URL", using the terminology from [4]. The page http://www.braces.org is generally found by the query "braces", arguably an informational query. However, people in the highest quintile of per-capita income find it predominantly by the query "braces.org", a navigational query. These are some anecdotal examples illustrating differences in how people search for the *same* web page. Table 6 shows that using demographic information can lead to an improvement in automatically deriving the query for a given URLs of 4.8% in terms of the precision for the highest ranked query.

6.3 Query Completion

If a user starts typing the word "frontpage" on a search engine, what is the most likely term to follow? For most people, it is "2003", referring to a particular version of the Microsoft software. However, for young people it is "free", for African Americans it is "africa"[14], and for people with a high level of education it is "magazine"[15]. In this set of experiments, we evaluated if demographic information could help in predicting the second query term. We used queries having at least two whitespace-separated terms, and having at least two characters in each of the first two terms.

This application shows the largest gain in Table 6. This make intuitive sense as the *full* query or URL already absorbs a large amount of demographic information. So the earlier the demographic information is used in the process, the larger its predictive value.

Recall that the percentiles used to discretize the demographic features, as described in Section 3.2, were computed for (query,URL) pairs for the web search setting. So the performance for the other two applications can most likely be improved further by adapting the discretization as this leads to a higher entropy $H(D)$ and hence to a more discriminative predictor.

7. FUTURE EXTENSIONS

Our findings in Section 5 are more *descriptive* than *explanatory*. It would be interesting to investigate why, for example, different demographic groups prefer different result URLs for the very frequent query "swine flu symptoms", where people in the lowest quintile concerning a BA degree have http://www.medicinenet.com/swine_flu/article.htm as the single most clicked result as opposed to http://www.cdc.gov/h1n1flu/qa.htm for the rest of the population, even though both pages appear to be of a similar nature. Apart from doing a questionnaire-based user study, one could look at features of the target page, such as the number of images or the average sentence length, or of the result snippet, such as which words were shown. Similarly, features such as the length of a session, the dwell time on a target page or even the percentage of typographical mistakes in queries could help to improve user modeling. We hope that a better understanding of the differences in how certain demographic groups interact with the web leads to more targeted help for people in need. E.g., one could envision a different web search interface for the elderly.

As our query log data seems to be representative of the whole US population (see Table 2), it lends itself for various large-scale sociological studies. For instance, it might be possible to investigate different attitudes of different demographic groups, such as women vs. men, towards certain off-line issues, e.g. child care, or online issues such as digital privacy concerns [24]. This could be done using both the generated search volume and the type of related search queries for the issue of interest. For instance, we observed that men appear to be more worried about deleting their search history while women tend to be more worried about removing their Facebook profiles.[16] Any study using query logs would be faster, cheaper and have wider coverage than traditional field studies. As a downside there is, on the other hand, more noise and effects due to spam or a small set of highly biased users have to be taken into account for such studies.

We also deem it interesting to investigate algorithms such as HITS [16] to propagate demographic labels back and forth between, say, users and queries. This way a user who often issues "educated" queries would gradually be labeled as more and more educated, and queries issued by very educated users would similarly become labeled as more educated. Of course, this approach could be applied to any of the demographic features considered.

Finally, we are looking at a possibility to share our data on a per-query basis for high-volume queries if privacy guarantees such as k-anonymity can be given [25]. Releasing query log information aggregated for demographic groups is similar in spirit to releasing census information for a particular ZIP code.

8. CONCLUSIONS

To the best of our knowledge, this is the first study that analyzes the web search behavior of different demographic groups, such as different income ranges or different ethnic groups, for millions of US web users. The simple but important observation that made this possible was the linkage of census information for ZIP codes to user profiles. For most parts, the population of search engine users appears to be a very good approximation of the US population (see Table 2), which highlights the potential of our approach for sociological studies.

It should be emphasized that we could compile lists of queries of people of Asian decent (Table 3) or point at differences in web search behavior for more educated people (Table 4) without knowing the ethnicity or the education level of a single person. We see this as a big advantage of our approach over more traditional questionnaire based field

[14] http://www.frontpageafrica.com/ is an Africa-centered news and entertainment site.

[15] http://frontpagemag.com/ is a political, conservative magazine.

[16] For the query "how to delete permanently" the most-clicked URL for women was http://www.ehow.com/how_2315204_delete-facebook-account-permanently.html while for men it was http://www.metacafe.com/watch/1267808/how_to_permanently_delete_google_search_history/.

studies, where demographic profiles need to be collected and stored for individuals, raising privacy concerns.

With respect to the impact of our approach on the web search results of a major web search engine, we demonstrated that a straightforward application of the demographic information led to a 1.4% increase in P@1 averaged among all searches, and an increase of 7.1% in P@1 for the 30% of queries having the larger entropy in their click distribution. See Section 6 for details. Though this might seem small, one has to take into account that (i) the baseline system is already highly optimized and that (ii) the full query itself already "absorbs" part of the demographic information, as certain user groups are less likely to issue certain queries. Bigger gains are achievable for query completions as here the demographic information is not yet subsumed by the full query. Changes to interfaces of this kind also have the advantage that users are less likely to be confused by the fact that the ranking of results is different between, say, friends.

Although in principle our methodology is applicable to any country, our current study is limited to the US due to the availability of detailed government census information. In cases where this information is not readily available, applying machine learning techniques seems viable.

Our main purpose was to point out opportunities which arise from using demographic information and we believe that we have barely scratched the surface. We hope to do but also to see more work on this topic in the future.

Key references: [27, 19]

9. REFERENCES

[1] S. Argamon, M. Koppel, and G. Avneri. Routing documents according to style. In *First International Workshop on Innovative Information Systems*, 1998.

[2] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni. Gender, genre, and writing style in formal written texts. *Text*, 23, 2003.

[3] C. Arndt. *Information Measures: Information and its description in Science and Engineering*. Springer, 2001.

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

[5] T. Coenen. Pink truth. `http://www.pinktruth.com`.

[6] M. Eirinaki and M. Vazirgiannis. Web mining for web personalization. *ACM TOIT*, 3(1):1–27, 2003.

[7] N. Ford, D. Miller, and N. Moss. The role of individual differences in Internet searching: An empirical study. *JASIST*, 52(12):1049–1066, 2001.

[8] N. Ford, D. Miller, and N. Moss. Web search strategies and human individual differences: A combined analysis. *JASIST*, 56(7):757–764, 2005.

[9] N. Ford, D. Miller, and N. Moss. Web search strategies and human individual differences: Cognitive and demographic factors, Internet attitudes, and approaches. *JASIST*, 56(7):741–756, 2005.

[10] D. L. Hoffman and T. P. Novak. Bridging the racial divide on the internet. *Science*, 280:390–391, 1998.

[11] I. Hsieh-Yee. Research on Web search behavior. *Library and Information Science Research*, 23(2):167–185, 2001.

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

[13] L. A. Jackson, K. S. Ervin, P. D. Gardner, and N. Schmitt. Gender and the internet: Women communicating and men searching. *Sex Roles*, 44(5):363–379, 2001.

[14] R. Joiner, J. Gavin, J. Duffield, M. Brosnan, C. Crook, A. Durndell, P. Maras, J. Miller, A. J. Scott, and P. Lovatt. Gender, internet identification, and internet anxiety: correlates of internet use. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, 8(4):371–378, 2005.

[15] R. Jones, R. Kumar, B. Pang, and A. Tomkins. "I know what you did last summer": query logs and user privacy. In *CIKM*, pages 909–914, 2007.

[16] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *JACM*, 46(5):604–632, 1999.

[17] L. Lorigo, B. Pan, H. Hembrooke, T. Joachims, L. Granka, and G. Gay. The influence of task and gender on search and evaluation behavior using google. *Information Processing & Management*, 42(4):1123–1131, 2006.

[18] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. In *Collaborative Web Tagging Workshop*, Edinburgh, Scotland, 2006.

[19] Q. Mei and K. Church. Entropy of search logs: how hard is search? with personalization? with backoff? In *WSDM*, pages 45–54, 2008.

[20] A. J. Morgan and E. M. Trauth. *Impact of Individual Differences on Web Searching Performance: Issues for Design and the Digital Divide*, chapter ITB12097, pages 261–282. Idea Group Publishing, 2006.

[21] M. R. Morris, J. Teevan, and S. Bush. Enhancing collaborative web search with personalization: groupization, smart splitting, and group hit-highlighting. In *CSCW*, pages 481–484, 2008.

[22] P. Norris. *Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide*. Cambridge University Press, 2001.

[23] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 2006.

[24] K. Sheehan. An investigation of gender differences in on-line privacy concerns and resultant behaviors. *Journal of Direct Marketing*, 13(4):24–38, 2000.

[25] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[26] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR*, pages 449–456, 2005.

[27] J. Teevan, M. R. Morris, and S. Bush. Discovering and using groups to improve personalized search. In *WSDM*, pages 15–24, 2009.