The Demographics of Mail Search and their Application to Query Suggestion

David Carmel, Liane Lewin-Eytan, Alex Libov, Yoelle Maarek, Ariel Raviv
Yahoo Research
Park MATAM, Haifa 31905, Israel
{dcarmel, liane, alibov}@yahoo-inc.com, yoelle@yahoo.com,@yahoo-inc.com

ABSTRACT

Web mail search is an emerging topic, which has not been the object of as many studies as traditional Web search. In particular, little is known about the characteristics of mail searchers and of the queries they issue. We study here the characteristics of Web mail searchers, and explore how demographic signals such as location, age, gender, and inferred income, influence their search behavior. We try to understand for instance, whether women exhibit different mail search patterns than men, or whether senior people formulate more precise queries than younger people. We compare our results, obtained from the analysis of a Yahoo Web mail search query log, to similar work conducted in Web and Twitter search. In addition, we demonstrate the value of the user's personal query log, as well as of the global query log and of the demographic signals, in a key search task: dynamic query auto-completion. We discuss how going beyond users' personal query logs (their search history) significantly improves the quality of suggestions, in spite of the fact that a user's mailbox is perceived as being highly personal. In particular, we note the striking value of demographic features for queries relating to companies/organizations, thus verifying our assumption that query completion benefits from leveraging queries issued by "people like me". We believe that demographics and other such global features can be leveraged in other mail applications, and hope that this work is a first step in this direction.

1. INTRODUCTION

Popular Web search engines receive daily billions of queries and collect terabytes of usage data signals, which for more than a decade have provided great research opportunities for improving ranking models, learning user behavior, etc. [18]. In contrast, Web mail search has attracted much less attention from the research community, probably due to privacy concerns and to the lack of publicly available datasets[1].

Mail search is much closer to desktop search than to Web search, as noted in [7]. The indexed corpus is relatively small and strictly personal, rather than shared between users, and search retrieval is optimized for recall rather than for precision. Users, very much like in desktop search [13], want to "re-find" a message they remember having read, while Web searchers will rarely know what they might have missed. With such differences in corpora and in users' search expectations, we should expect different types of queries and search behavior. As a first evidence. We have verified that Web mail search queries have an average size of 1.5 terms and are thus half the size, again on average, than Web search queries as reported in [28].

Short queries are to be expected, as mail users optimize for recall: they prefer issuing a vague query and then exhaustively browse results, sorted by time, in the hope of identifying the message they seek. This highlights yet another fundamental difference between Web search, where results are ranked by relevance and mail search, where they are mostly ranked by time, even if there have been recently some efforts to introduce relevance ranking in Mail search [7].

In this work, we propose to study the nature of mail search queries at three different levels of granularity (1) individual, considering only the user's personal query log (2) global, as provided by the entire US mail searchers population and finally (3) demographic, considering logs provided by users sharing the same location (as defined by the state), age, gender, and predicted income.

We propose to verify whether we can leverage the insights derived from these analyses, in order to improve one of the search mechanisms that does the most extensive use of query logs, namely query auto-completion. While this area has been thoroughly studied in Web search, it has been mostly ignored in Mail. The mail search query completion experience seems to be still evolving, with some mail search services suggesting only past queries from the same user, others offering contact names and message titles, or strings derived from message bodies. This task is doubly challenging as a user's inbox is highly personal and mail searchers issue much fewer queries than Web searchers.

Our conjecture here is that users will benefit from queries previously issued to mail search service by "people like them". In the same way that auto-completion in Web search differ by country, with "liv" being completed to "liverpool" in UK, as opposed to "liver" in the US[2], we would expect for example that the prefix "be" be completed to "berkeley" for a young student in California, as opposed to "best buy" for a middle-age person in Texas, even if none of them had ever issued this specific a mail search query in the past.

We present a supervised learning-to-rank framework for ranking query auto-completion candidates for mail search, by augmenting

[1]Note that we consider Web mail rather than enterprise mail, whose characteristics are different, and for which public datasets such as the Enron corpus [22] do exist.

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052658

[2]See "Local flavor for Google Suggest" in https://googleblog.blogspot.co.il/2009/03/local-flavor-for-google-suggest.html

features representing the individual user's query log with features extracted from the global query logs of millions of other users, while considering demographic similarities with the original user.

The key contributions of this work are twofold: (1) we present the first, to the best of our knowledge, large scale analysis of Web mail search query logs, as provided by Yahoo Web mail logs[3], specifically focusing on the demographics of mail searchers and (2) we demonstrate how global signals and demographic attributes can be leveraged in order to increase the quality of automatic query completion for mail search.

2. RELATED WORK

Mail search is an emerging field with much potential for wider research, but lacks representative public datasets due to its highly private nature. A few exceptions are the public *w3.org* mail data used by the TREC Enterprise Tracks [9, 25], and the Enron dataset [22], which opened the Enron corporate email data to the public more than a decade ago.

Conversely, the large scale data collected in Web search has been a fertile ground for substantial research, where users search behavior has been widely explored. For example, Spink *et al.* [26] analyzed the query log of the Excite Web search engine, studying web searchers' behavior in terms of query length, number of results, reformulations and usage of advanced search tools. Jones *et al.* [19] studied typical session behavior in Web search, while Teevan *et al.* [27] explored repeated queries in Yahoo's query log, revealing that about 40% of all queries are re-finding queries.

Weber *et al.* [31, 32] studied the demographics of Web Search on Yahoo search users. They analyzed various search behaviors across user demographics such as gender, age, income, and state. They showed how user demographics can contribute to query suggestions, as popular queries and user preferences may vary drastically across different demographics. Shokouhi [24] extended the demographics-based user model by taking users' long-term search history and their location into consideration. The contribution of the user search history for personalized query auto-completion has been demonstrated in many previous studies (e.g., [6, 23, 3]).

Teevan *et al.* [28] compared search behavior on Twitter and on the Web, observing that Twitter queries are shorter, more popular, and less likely to evolve than Web queries.

In the absence of a query log, several studies tried to predict the searcher demographics. Hu et al, [16] predicted users' gender and age from their Web browsing behaviors. Demographic prediction was also applied on Twitter traffic data [10], and on mobile data [33, 12]. Bi *et al.* [5] showed how user demographic traits such as age, gender, and even political and religious views, can be efficiently and accurately inferred based on the search history.

Demographic analysis has been widely applied for personalizing Web search ranking. Bennet *et al.* [4] personalized the ranking model based on the similarity of the locations of search results to the user locations. Kharitonov *et al.* [21] learned a context-aware relevance model from user clicks, showing that the demographic-based ranking features provide significant improvements in ranking quality. Similarly, demographic features were shown to contribute to ad-targeting in sponsored search [17, 30].

In this work, we follow the same direction as Weber *et al.* [31, 32] from 2010-2011 and Teevan *et al.* [28] from 2011, in the context of mail search, which has remained unexplored until today. We compare our findings to theirs in order to better understand the effect of demographic attributes on users' mail search.

[3]We emphasize that all the experiments reported here have been conducted on fully anonymized data, in accordance with Yahoo strict privacy policy.

3. QUERY LOG ANALYSIS

The main source of information used for our study is the mail search query log of Yahoo Web mail, limited to US mail search traffic. The data was collected over one month (from March 18 to April 18, 2016), and includes mail search queries, result page information, and following click events related to the search results. Over this period, we used a sample of about 50M queries and 63M click events, for 5.5M US mail users. To protect users privacy, users' identifiers were hashed using a non-invertible function.

3.1 Demographics

We used the profile information provided by registered users, the main attributes being: gender, age and birth year. We also obtained state and ZIP code information based on the users' IP. Additionally, we joined the ZIP code information with publicly accessible demographic information obtained from US-census[4] in order to derive a median income based on the user's ZIP code, following the methodology[5] of Weber et al. in [31]. Table 1 summarizes the per-query as well as per-user demographics of our dataset, and compares them with averages of the US population (obtained from US census) and with the Web analysis presented in [31].

Demog.	Average per-query		Average	US
	mail	Web	per searcher	avg.
Age	46.45	41.3	45.98	37.7
Income (K)	32.83	22.7	32.85	28.88
Male	41.63%	50.3%	44.96%	49.2%
Female	58.37%	49.7%	55.04%	50.8%

Table 1: Demographics per query and per user, based on our query-log data, compared to the US average from census data.

Comparing our numbers to the demographic analysis of Web search from [31], we can see sizable age and income differences (average age of 41.3 and average income of 22.7K in Web vs. 45.98 and 32.85K in mail respectively). The age difference implies that mail searchers are an older population compared to Web searchers. This suggests that the younger generation makes less usage of mail as it is more engaged in social and messaging platforms as remarked by [29]. There is also a noticeable difference between average income of mail searchers and Web searchers, probably due to their older age as well as the inflation since the 2010 measurements ([31]). We note that the demographic trends presented in this section are specific to the Yahoo Mail user base.

3.2 Analysis

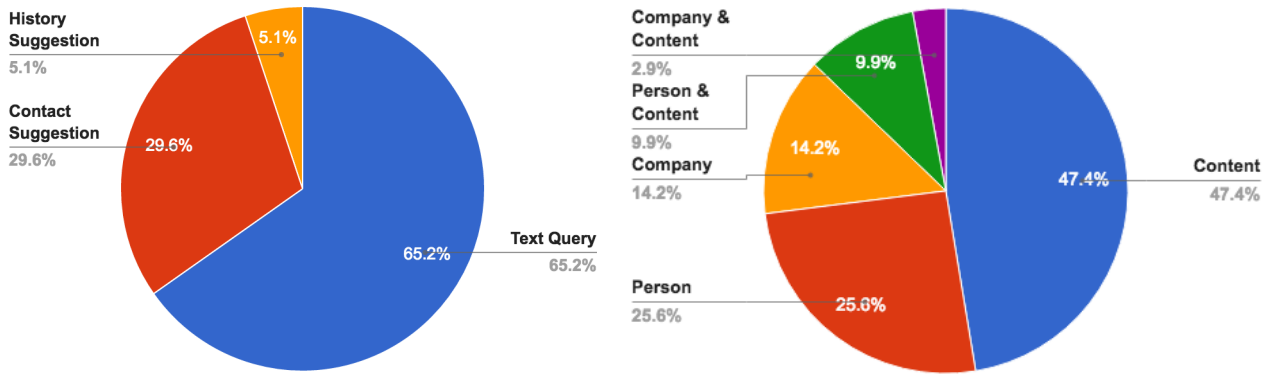
We analyze different characteristics of mail search, focusing mainly on queries and on the quality of their results. We examine the entire sample population (US) as well as different demographic sectors.

3.2.1 Query Structural Analysis.

Query Source. We focus on the source of the query, and consider the two sources from which a query can originate: *text* queries that

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

[5]This methodology is clearly a very coarse approximation, yet since it has already been used in the context of Web search, it serves our purpose of comparing mail and Web search demographics.



(a) Distribution of query source

(b) Type distribution of text queries

Figure 1: Distributions of query source and query type

Demographics	Text queries%	Contact suggest.%	History suggest.%
Male	66	28.9	5.1
Female	64.8	30.1	5.1
Age (0-20)	64.3	28.5	7.2
Age (60-80)	58.8	34.4	6.8
Income (0-20%)	65.2	29.2	5.6
Income (80-100%)	66.8	28.7	4.5
California)	66.7	28.9	4.5
Mississippi	61.5	31.7	6.7
New York	66	29.1	4.9
All (US)	**65.2**	**29.6**	**5.1**

Table 2: Query source distribution.

are fully formulated by the user, and *suggestions*, i.e., queries that were selected by the user from a suggestion list. Suggestions are further divided into two types (1) *contact suggestions*, based on the user's contact list, and (2) *history suggestions*, based on the user's mail search history. Both are dynamically presented to users as they type. We observed that out of all queries submitted to the system, 65.2% are fully formulated by the user, 29.6% consist of contact suggestions, while 5.1% are historical suggestions.

Figure 1(a) presents the distribution of the query source for our sample, and Table 2 presents a similar distribution for the different demographic subsets of users.

Query Type. We analyze the two different query types identified in mail search. The first are *content* queries, which include textual strings that the user attempts to match to terms in the content of the message, or in its subject (or, possibly in the content of a document attached to the message). The second are referred to as *contact* queries, which include names or addresses of contacts (sender, for incoming mail, or receiver, for outgoing mail).

More specifically, when focusing on contacts, we further identify two sub-types. The first, referred to as *person queries*, includes names or mail addresses of persons. The second, referred to as *company queries*, includes names of organizations or companies (e.g. "amazon", "southwest", "facebook"). Person and company queries clearly have a similar intent of identifying messages sent to or received by a specific person or company.

Both person and company queries can originate from suggestions, as well as be fully formulated by users. However, suggestions of companies as contacts are not frequent, as they are less likely to be part of a user's contact list. Contact suggestions are identified as representing 29.6% of the queries in our sample. Contacts and companies can also be found in text queries fully formulated by users. Thus, we further identify persons and companies in the text queries (65.2% of all queries) by matching the query strings against dictionaries of names and top domains of companies and organizations (ranked according to traffic volume). Figure 1(b) illustrates the overall distribution of query types.

Overall, we observe a high percentage of contact queries (about 55%, not including queries with additional terms). In fact, such queries simply serve as filters, for cases in which the user is interested only in messages from/to a specific contact, and then exhaustively browse the results (typically ranked in reverse chronological order) in order to find the wanted message.

In the rest of Section 3, the analysis considers only text queries that are fully formulated by the user.

3.2.2 *Query Lexical Analysis*

Query Length. First, we consider query length with respect to the number of terms and the number of characters in the query. Mail search queries are very short: almost 70% of the queries are single term queries and about 20% of the queries consist of two terms, with an average of 1.49 terms for the entire dataset, as mentioned earlier. We argue here that such short queries are typical of users who struggle with issuing a well defined query and try to increase recall. A short query allows users to get back a large set of results, which are, in most clients, ranked by time, and then exhaustively browse the results list, often using the timestamp as an additional signal, in order to make sure they are not missing the relevant one(s). Table 4 lists the average mail search query lengths for the different demographic subsets of mail users as compared to the entire US traffic.

Frequent Queries. We investigate the lexical attributes of mail search queries and present the most frequent unigram, bigram and trigram queries for the entire sample in Table 3. We can see that across query lengths, the most popular queries are mainly names of companies and organizations. This follows the new nature of Web mail traffic, in which more than 90% of non-spam Web email is generated by automated scripts that send messages on behalf of a company or an organization (examples are shipment notifications, flight itineraries, social events, monthly bank statements, etc.), [1, 14].

Queries Language Models. Next, we focus on the differences between the language models across demographic groups by examining the most "discriminating" queries for each group, as compared

Unigram Queries	Bigram Queries	Trigram Queries
amazon	turbo tax	bank of america
resume	american airlines	shop your way
southwest	home depot	the home depot
facebook	best buy	the childrens place
groupon	capital one	southwest airlines co
ebay	wells fargo	sign up genius
tax	old navy	save the date
kohls	state farm	time warner cable

Table 3: Most frequent unigram, bigram and trigram queries.

Demographics	Query Length by Chars	Query Length by Terms
Male	9.53	1.47
Female	9.76	1.51
Age (0-20)	8.82	1.41
Age (60-80)	10.98	1.66
Income (0-20%)	9.61	1.48
Income (80%-100)	9.53	1.48
California)	9.58	1.48
Mississippi	10.1	1.55
New York	9.61	1.48
All (US)	**9.67**	**1.49**

Table 4: Query length measures

to the rest of the population. Table 5 lists five examples out of the top-20 discriminating queries for different demographic groups, along their *Kullback-Leibler divergence*(KL) values.

Figures 2 illustrates discriminating queries in side-by-side word clouds, where the more discriminating the term, the larger the font.

This analysis is aligned with common stereotypes, however as mentioned before, it reflects the probabilities inferred from the millions of queries and users. Looking at the top discriminating queries for the different demographic groups, we observe the following: Men are more identified with technology, online-gaming and finance than women, who in turn search more often for messages from large retailers and social media domains, as compared to men, thus addressing one of the questions formulated in the Introduction. In addition, education-related queries are indicative of the younger population, while names of home-shopping networks are more prevalent in searches of the elders. Finally, we note that airlines and hotels names are more often associated with the upper class, whereas queries related to job-seeking and social media are more indicative of the lower class. Interestingly, when comparing these results to their Web counterparts as reported by Weber et al.[32], we see "stereotypical" similarities such as men's interest in technology and women's in shopping, but also some gaps, e.g., the lack of sports-related queries in the mail domain.

[6]Digital distribution platform for multiplayer gaming
[7]Free Application for Federal Student Aid
[8]College admissions test
[9]Online sweepstakes and shopping site
[10]American Association of Retired Persons
[11]Broadcasting network specializing in televised home shopping
[12]Home Shopping Network

Demographics	Query	Value ($*10^{-4}$)
Gender Male	newegg	3.00
	tax	2.60
	dell	2.10
	turbotax	1.80
	steam[6]	1.54
Gender Female	kohls	10.31
	target	5.84
	macys	5.42
	facebook	4.56
	pinterest	2.92
Age "young" (0-20)	fafsa[7]	212.58
	ucla	141.56
	act[8]	89.53
	cornell	76.63
	scholarship	61.71
Age "senior" (60-80)	google	14.83
	pch[9]	5.60
	aarp[10]	5.43
	qvc[11]	4.63
	hsn[12]	3.07
Wealth Lower Class (0-20%)	resume	14.23
	fafsa	3.77
	facebook	3.71
	walmart	2.87
	amazon	1.97
Wealth Upper Class (80%-100)	evite	4.50
	united	3.27
	southwest	1.92
	donation	1.30
	marriott	0.89

Table 5: Highly discriminating queries for different demographics along their *Kullback-Leibler divergence* (KL) values.

3.2.3 Query Results.

We analyze the quality of the query results according to different measures, both for the entire sample as well as for the demographic subsets. Following common practice [19], we treat queries that occur in a sequence with no inactivity interval of 15 minutes length to be part of the same session, and consider the results for the last query of every session.

As a simplifying assumption, we consider a query successful when the user clicked on at least one of the search results. The *Success rate* is the percentage of successful queries. Queries not followed by any click are divided into two disjoint groups: queries with no results (no messages matched the query), and queries with no click (no result was clicked following the query). We note that the last scenario does not always correspond to a failure, as it could be that users find the needed information in the message snippet, yet this case is most probably much less frequent than in Web search where Direct Displays are more informative [8]. We report our findings in Table 6.

3.3 Discussion

We infer some interesting (yet admittedly often stereotypical) insights from our various analyses.

Gender. Women generally type longer queries than men (on average, 1.51 and 1.47 terms respectively), and engage in longer sessions (2.2 compared to 2.0, where the session is measured in terms of number of queries). We note that longer queries may extend

Figure 2: Word clouds of discriminating mail search queries for men (on the left) vs women (on the right)

Demographics	Failure		Success
	No Results %	No Clicks %	Clicks %
Male	9.4	21.3	69.3
Female	9.1	21.5	69.4
Age (0-20)	7.8	19.6	72.6
Age (60-80)	14.3	26.0	59.7
Income (0-20%)	9.5	21.8	68.7
Income (80-100)	8.2	20.5	71.3
California	8.7	20.3	71.0
Mississippi	11.0	23.7	65.3
New York	8.0	20.8	71.2
All (US)	**9.3**	**21.5**	**69.2**

Table 6: Results quality for the different demographic subsets

sessions due to the way results are matched against the query, following the hard constraints imposed by the traditional time ranking used in mail search [7], where typically all query terms must appear in the message. Interestingly, women click on people suggestions more often than men.

Age. Young people write much shorter queries (8.82 characters) and have a short session length (1.98), but still manage to get a high Success rate (72.6%). This suggests that young people choose correct distinctive terms to get their desired results. Conversely, senior people write much longer queries (10.98 characters) and often abandon the search session without selecting any result (26%) or without getting results at all (14.3%). This highlights their potential benefit from being presented more advanced personalized search suggestions.

Income. Users in the lowest income percentiles, as inferred from their ZIP code, exhibit a higher failure rate as compared to higher percentiles (31.3% and 28.7% respectively), while their queries length is similar. This is likely rooted in poor query formulation of this demographic sector.

Location. California and New York measurements are correlated with the numbers of rich/young population, while Mississippi is correlated with the opposite demographic segments.

3.4 Comparison to Other Domains

Several attributes arising from our analysis are clearly different from parallel attributes in Web search, following the different nature and usage of these two domains. It is interesting not only to see how mail search behavior differs from Web search behavior, but also to compare with search in social networks. Table 7 compares some of our findings to those reported in the Web [31] in 2010 and on Twitter [28, 20] in 2011 and 2013 respectively. First, we note that the average number of query terms in mail (1.49) is shorter than in the other domains. Interestingly, it is much closer to the average on Twitter, (1.64) than on the Web (3.08). Short queries in mail are most probably an effect of the users attempting to "shoot wide" in order not to miss the relevant message they are seeking, as well as of the the way results are matched and ranked, as mentioned before. We observe the same pattern when examining the differences in session length (measured as the number of queries that occur in a sequence with no inactivity interval more than 15 minutes). The difference in session length is likely a result of the user's search intent: trying to re-find familiar data in mail, versus seeking new information in the Web.

Next, we explore the differences in the average ratio of repeated queries in all three domains, measured by evaluating the average ratio of non-unique queries each user issues over the entire dataset. Evidently, users repeat their queries in mail more frequently than in the Web (45% and 34.7% respectively) and even more often in social networks (55.8%). The large gap can be attributed to a different personalization level of the content in these domains: from purely personal in mail to public information in Web. Repeated queries in social networks, which are in between personal and public, are often the result of the user's trying to access renewed information from the same origin as per [28].

Finally, we consider the click entropy in the three domains, represented by the distribution of user's clicks on different positions in the results list. We note that the click entropy in mail, 2.95, is comparable to that in Twitter, and almost twice the value of click entropy in the Web ($1.60 - 1.74$). This is probably a result of the way search results are ranked. While Web search engines generally use highly sophisticated methods to rank results by relevance, offering the user the most useful results in top positions, mail services, as well as Twitter, typically rank search results by recency, forcing the user to browse deeper in order to find old yet relevant results.

4. MAIL QUERY SUGGESTION

Given the differences in search behavior across various demographic groups, as reported in the previous section, we decided to further investigate the influence of demographic attributes on a specific mail application: query suggestion. This application is a natural candidate for this validation exercise, as it heavily relies on query logs. Moreover, given that mail search queries are on average very short and non specific, as detailed in Section 3, we believe that assisting users in formulating longer and more specific queries will significantly improve their mail search experience.

More specifically, we investigate the contributions of global signals and the demographic properties introduced in Section 3, on the query completion task, as compared to the user's personal query log, which currently serves as the main source for query completion in Yahoo Web mail.

In the most common settings, the query suggestion mechanism takes as input a short string of characters entered by the user (the *prefix*) and returns a ranked list of fully formulated queries (the *suggestions*). For the sake of simplicity, we only consider here completions of the prefix entered by the user, using actual queries originating from query logs, but note that other methods have been proposed to derive suggestions using the indexed corpus or other methods [15]. We do not consider the influence of the actual mes-

	Mail	Twitter	Web
Query length (by Terms)	1.49	1.64	3.08
Session length (by Queries)	2.12	2.20	2.88
Repeated queries (%)	45.0	55.8	34.7
Click Entropy	2.95	2.93-4.13	1.60-1.74

Table 7: Mail search statistics compared to Twitter and Web

sage content, which we reserve for future work. The ranking of suggestions can be done via various methods like in other information retrieval tasks, from occurrence frequency considerations to learning-to-rank methods, which we chose to adopt here. Query completion methods have been studied in details in the context of the Web, where query logs common to all users are huge and a wisdom of crowd approach is highly beneficial, see Bar-Yossef *et al.* [3], as an early example of such studies. However in the context of mail search, the task is more challenging as detailed below.

4.1 Personal vs Global Suggestions

One of the reasons of the success of query completion in Web search is that suggestions originate from a huge query log that is common to millions if not hundreds of millions of users. In contrast, the mail domain is personal, so the default approach has been to leverage only the individual user's query log as source of suggestions. The major drawback of this approach is that, in Web mail, most users initiate on average only a few queries per month. With such small query logs, the query suggestion mechanism achieves low coverage. We therefore propose to challenge the default assumption that other user's inboxes are useless as a whole. We study what level of query log generalization, from the country level, as done in Web search[3], to at a finer grained level, e.g. using demographic attributes as presented in Section 3, might bring value. Our intuition here is that even if one's mailbox is highly personal, the machine-generated messages, so dominant in today's Web mail, are common to many "similar users", who should have similar needs and might benefit from each other's queries. As a first validation, we verified that there exist head queries shared by large numbers of users, as seen in Table 3, most of these queries refer to commercial entities or organizations that lead to machine-generated messages. This approach of augmenting the suggestion corpus by considering "similar" users, where the notion of similarity needs to be defined, follows Baeza-Yates *et al.*'s intuition that "employing wisdom of crowds to bias the results of a query is only worth if the users share the crowd's values" [2]. We note that global suggestions should be further validated to match the content in the user's mailbox in order to guarantee a non-zero result set, however this is out of the scope of this paper.

We evaluate below the personal vs. global features in the query completion task, and compare the contributions of demographic-driven features.

4.2 Ranking Completion Suggestions

We have chosen to apply here a learning-to-rank approach, in order to select the best query completion suggestions. We selected, AROW, an online variant of SVMRank, which learns a linear weight vector through pairwise comparisons between the relevant candidate and other top-ranked candidates for each query. We tuned the learning parameters through standard training and validation on separate sets, while testing on a different set.

Note that in the mail domain, generating a training set is challenging as the data is private and sensitive. One cannot use external evaluators for manual labeling as personal mailboxes cannot be released externally for obvious privacy reasons. In any case, mailboxes are so personal, that it is questionable that anyone but the owner of the mailbox would be truly capable of adequately inferring the intent of the query and selecting the right result. We therefore followed Shokouhi [24] and employed a method for automatically generating labels. First, we sampled a set of queries followed by a clicked result from our query log and decomposed each query into the set of all its prefixes. For each prefix, we obtained all matching query candidates using an auto-completion trie structure over the entire sample. Next, we marked the candidate that is identical to the query submitted by the user as relevant, for each prefix. We considered only queries that were followed by a click on a result in order to avoid low-quality queries, such as misspelled or malformed queries.

Next we list the features we used for generating suggestions. We divide the features into three types as detailed below:

Personal user features: These features are simply based on the user's personal query log. We use the log likelihood function over the raw historical frequencies computed from the background data.

Global (US) features: These features are based on the users' global query log, considering the entire US mail searchers population. As for the personal features, we use the log likelihood function over the raw historical frequencies computed from the background data.

Demographic features: We consider four basic families of features: location, age, gender and inferred income, each corresponding to a different demographic group. We derive for each suggestion candidate its past frequency within each group. We apply Laplace smoothing over the raw probability to cope with data sparsity issues, and use the log likelihood score as a feature in our framework, in order to accommodate for the linear properties of our LTR model. The demographic features are computed based on the user profile information. The age-based features are computed by splitting users into five age groups {below 20, 21-40, 40-60, 60-80, and above 80}. Location features are based on the user's state of residence (derived from their ZIP code in the US). Finally, the inferred income features are inferred by joining the user's ZIP code information with US-census as described in Section 3. We split the users into five groups, according to their per capita income percentile, i.e {below 20, 21-40, 40-60, 60-80, and above 80}.

4.3 Experiments

4.3.1 Dataset

We sampled the query log for 23.7M queries issued by 4.4M unique users between March 21, 2016 and April 16, 2016. Following Shokouhi [24], we filtered out queries that appeared less than 10 times as these are often malformed or too rare to be highly ranked in auto-completion lists. In order to conform with data privacy and security constraints of the mail domain, we also removed numbers and special characters from the query candidates in a preprocessing step, thereby stripping suggestions of account numbers, addresses and other personal information, using k-anonymization methods similar to those used in recent work by DiCastro *et. al* [11].

The entire dataset was partitioned as follows. We took from the above dataset a subset of 20.8M queries issued until April 14, 2016 as background data for generating the tries per demographic sections and forming the search history of users. In addition, for each prefix length, we took 20K queries issued between April 15 and April 16, 2016 to form a validation set, which was used to tune a linear model that combines the different features using an LTR approach, as explained earlier. In a similar manner, for each prefix length, we selected 20K queries issued between April 17 and April 18, 2016 in order to form a test set. For each prefix length, we collected only queries of longer length, to avoid the trivial case where the relevant candidate is the prefix itself. This way, we guarantee that all sets are non-overlapping and that we remain close to real-world settings.

In all our experiments, we use the same set of candidates per prefix, in order to highlight the improvement achieved by re-ranking. The initial set of suggestions provided as input to our ranker consists of the top-100 candidates, selected by assigning uniform weights to all features. We use the Mean Reciprocal Rank (MRR) of the rel-

evant suggestion as quality measure, as there is only one relevant candidate per prefix for each query, the one the user eventually submitted. We also use the Success@1 metric to highlight the cases where the top ranked suggestion was indeed the relevant one. The performance is measured with respect to different prefix lengths: $2, 3, 4$ characters and first term. Intuitively, the prefix length should have a significant influence: for a short prefix the number of candidates might be very large, thus the potential gain from re-ranking is higher, while for a long prefix the initial candidate list is limited as the user intention gets clearer. We also examine the performance of each feature group separately and provide a fine-grained breakdown of the demographic features and their effectiveness. Note that, given the nature of offline dataset, our evaluation is conservative: a suggestion that might have been relevant but differs from the one the user issued will be considered as a non relevant result.

4.3.2 *Combining Personal, Global, and Demographic Features*

In order to analyze the contributions of the personal, global and demographic-driven features, we conducted several experiments, adding the various types of features one at a time and then conducted an ablation test to evaluate the core contributions of demographic features. In addition, we compare the contribution of the demographic features to the global query log based on the whole US traffic.

Table 8 presents the effectiveness of our auto-completion method in terms of MRR, with the best results being achieved when combining personal and demographic features. First, it is worthwhile to note the influence of the prefix length: when considering only personal features, the MRR score increases from 0.276 for two characters, to 0.365 for four characters and to 0.441 given the first term, as listed in the "Person." column. The "Person.+US" column presents the results obtained by using features pertaining to the personal user's query log as well as features derived from queries of the whole US traffic, and the "Person.+Demog." column shows the results obtained when leveraging all demographic features. It can be seen that the improvement in MRR obtained by combining personal and demographic features is the better of the three cases, with an improvement in MRR between 60% to 114% depending on the length of the prefix. Looking at the Success@1 results, we see a similar influence of the features, but with an even higher increase (65%-125% depending on the prefix length). We note that we achieve an MRR increase of 2%-3% when taking into account the demographic features rather than considering the entire US traffic indiscriminately. In addition, adding global US features in addition to personal and demographics features, brings no improvement, most probably because the location feature (represented by the State) achieves the same result. Note that all reported gains, noted as (+xx%) in this Table and the following ones, are statistically significant, as validated by a two-tailed paired t-test ($p < 0.05$).

We provide a breakdown of the demographic features and analyze their effectiveness in terms of MRR gains. We focus our analysis on the results obtained for a prefix length of three characters as the performance of other lengths exhibit similar properties. Our results are presented in Table 9. We note that the inferred income feature, unlike the others, does not contribute to an increase in performance, in spite of the discriminating results we presented in Section 3. A possible reason could be that the way we inferred the income levels is too coarse-grained for this task, or simply that its influence is negligible considering other features.

Table 10 presents an ablation test we conducted in order to measure the core impact of demographic features. The first column gives the MRR per prefix length, considering, for comparison purposes, features of the entire US traffic, while the second column shows the results obtained using exclusively demographic features.

We observe that the demographic features per se provide an improvement ranging between 2%-4% in MRR and 2%-5% in Success@1 as compared to using US-based features. We also conducted an ablation test showing the contributions of personal features (see the second column of Table 8). It can be seen that the demographics features alone systematically contribute more than personal or global US features, yet the best results are achieved by combining personal and demographic features. Adding US features to all other features has no positive impact on either MRR or Success@1 most probably because their being redundant with demographic features.

	Prefix Length	Person.	Person. + US	Person. + Demog.
MRR	2 chars	0.276	0.479 (+73.6%)	**0.486 (+76.4%)**
	3 chars	0.279	0.588 (+110.7%)	**0.597 (+113.8%)**
	4 chars	0.365	0.714 (+95.5%)	**0.721 (+97.6%)**
	1 term	0.441	0.699 (+58.5%)	**0.709 (+60.9%)**
Succ@1	2 chars	0.231	0.377 (+63.0%)	**0.382 (+65.2%)**
	3 chars	0.221	0.476 (+115.2%)	**0.483 (+118.6%)**
	4 chars	0.274	0.607 (+121.6%)	**0.615 (+124.2%)**
	1 term	0.350	0.605 (+72.7%)	**0.615 (+75.7%)**

Table 8: The effect of combining personal, global (US), and demographic features on MRR and Success@1 for different prefix lengths

Features	MRR
Personal	0.279
Personal+Gender	0.582 (+108.3%)
Personal+Gender+Age	0.592 (+112.1%)
Personal+Gender+Age+Wealth	0.591 (+111.8%)
Personal+Gender+Age+Wealth+Location	0.597 (+113.8%)

Table 9: Breakdown of the demographic feature group effectiveness in terms of MRR for a prefix length of 3 characters. Gains are compared to personal features only.

	Prefix Length	US	Demographics
MRR	2 chars	0.300	**0.312 (+4.2%)**
	3 chars	0.466	**0.480 (+2.9%)**
	4 chars	0.628	**0.640 (+1.9%)**
	1 term	0.555	**0.574 (+3.4%)**
Succ@1	2 chars	0.192	**0.202 (+5.0%)**
	3 chars	0.339	**0.350 (+3.4%)**
	4 chars	0.506	**0.517 (+2.3%)**
	1 term	0.440	**0.458 (+4.2%)**

Table 10: The contribution of demographic features only as compared to all US traffic, measured in MRR and Success@1 for different prefix lengths.

We further illustrate the effectiveness of our ranker by presenting two examples from our dataset, reflecting some of the discriminating queries presented in Section 3. In both examples, we list the top suggestions returned by the ranker, after getting the first two characters as input, with and without demographic features. The first example is a 26 years old man from California who submitted the query $blizzard$. Table 11 shows that the ranker based on the US traffic only returned $bloomingdales$ as the top suggestion, while the relevant candidate, $blizzard$, is found at the $9th$ position. However, taking into account the demographic features, the ranker boosts the relevant candidate and positions $blizzard$ as the top suggestion. The second example is a 48 years old woman from Florida who submitted the query $target$. Table 12 shows that the

ranker based on the US traffic only returned tax as the top suggestion, while the relevant candidate, $target$, is second. However, taking into account the demographic features, the ranker positions $target$ as the top suggestion.

Rank	US	Demographics
1	bloomingdales	**blizzard**
2	blue cross	block
3	block	blake
4	blake	bloomingdales
5	blue	blue shield
6	blue apron	blue cross
7	blair	blue
8	blinds	bls
9	**blizzard**	blue apron
10	bls	blinds

Table 11: Top-10 suggestions, with and without demographics, for prefix bl submitted by a 26 years old man from California.

Rank	US	Demographics
1	tax	**target**
2	**target**	tax
3	taxes	taxes
4	taxact	taxact
5	tax return	tammy

Table 12: Top-5 suggestions, with and without demographics, for prefix ta submitted by a 48 years old woman from Florida.

4.3.3 Features Contribution by Query Type

Query Type	Person.	Demog.	Person. + Demog.
Company	0.245	0.794	**0.819 (+234.4%)**
Person	0.280	0.479	**0.596 (+112.7%)**
Content	0.285	0.373	**0.505 (+77.5%)**
Company & Content	0.245	0.160	**0.356 (+23.9%)**
Person & Content	0.245	0.196	**0.519 (+25.0%)**

Table 13: MRR results for different query types, given a prefix of 3 characters. Gains are compared to Personal.

We complete our experiments by evaluating the contribution of features, per query type, where we identify query types using the same method detailed in Section 3.2.1. Table 13 presents the MRR results obtained when using personal features alone, demographic features alone and then combining both types of features. The benefits of demographic features is the highest for the “company" query type, where they, alone, achieve an MRR higher by 224% than personal features alone. An even higher MRR is achieved when combining both types of features, with an improvement of 234% in MRR. Significant improvements in person and content queries can also be observed when using both types of features (113% and 77% respectively). We note that for longer queries that consist of both a contact (a company or a person) and additional content terms, the contribution of the demographic features is lower (24%-25%). Intuitively, as longer queries express a more specific intent of the user, global suggestions become less beneficial. Company or organization types typically refer to mass senders of machine-generated messages shared by many users, and mostly by “people like me" who will shop with the same vendors or interact with the same organizations as discussed in Section 3. The striking difference in value of demographic features for company-type queries verifies our original assumption that search suggestions benefit from demographic-based suggestions, due to the high volume of machine-generated messages.

5. CONCLUSION

This is, to the best of our knowledge, the first study analyzing the characteristics of Web mail searchers, investigating signals originating from the user’s personal log, a global query log as well as demographic signals. We noted intriguing as well as stereotypical findings, that highlight the differences in mail search usage between different demographic groups, and compared our results with those achieved in other search domains. The fundamental difference between mail search and Web search was apparent throughout our analysis. Mail searchers issue much shorter queries, engage in shorter sessions, and repeat their queries more often than Web searchers. We also noted that mail search is closer in its characteristics to search in social networks.

In the second part of our work, we demonstrated the benefits of the personal, global and demographic signals, by leveraging them in a query auto-completion system that we specifically tailored to mail search. More specifically, we demonstrated, via offline experiments, that combining personal log features with global and demographic features (derived from query logs of “people like me") achieves the best results, with MRR scores higher by 60%-114% (depending on the length of the prefix) than the MRR scores obtained when using a personal query log alone. Following several ablation tests, we verified that the demographic features bring more value than other feature families. In addition, we analyzed the influence of these features with respect to different query types, and demonstrated that the highest contribution of demographic features is obtained for contact-company queries, which typically target machine-generated messages.

The next step left for future work, in the context of query completion, is to extend our model to include the content of the user’s mailbox. We note however that following the highly sensitive nature of this data, it is far from being trivial. Additionally, the effect of such signals on message ranking itself is yet to be explored. We hope these new insights will inspire others to further investigate demographic signals in mail, and consider them in existing or yet to be invented mail features.

6. REFERENCES

[1] N. Ailon, Z. S. Karnin, E. Liberty, and Y. Maarek. Threading machine generated email. In *Proceedings of WSDM*, New York, NY, USA, 2013. ACM.

[2] R. Baeza-Yates, P. Boldi, A. Bozzon, M. Brambilla, S. Ceri, and G. Pasi. *Trends in Search Interaction*, pages 26–32. Springer Berlin Heidelberg, 2011.

[3] Z. Bar-Yossef and N. Kraus. Context-sensitive query auto-completion. In *Proceedings of WWW*, New York, NY, USA, 2011. ACM.

[4] P. N. Bennett, F. Radlinski, R. W. White, and E. Yilmaz. Inferring and using location metadata to personalize web search. In *Proceedings of SIGIR*. ACM, 2011.

[5] B. Bi, M. Shokouhi, M. Kosinski, and T. Graepel. Inferring the demographics of search users: Social data meets search queries. In *Proceedings of WWW*. ACM, 2013.

[6] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li. Context-aware query suggestion by mining click-through and session data. In *Proceedings of KDD*. ACM, 2008.

[7] D. Carmel, G. Halawi, L. Lewin-Eytan, Y. Maarek, and A. Raviv. Rank by time or by relevance?: Revisiting email search. In *Proceedings of CIKM*. ACM, 2015.

[8] C. Castillo, A. Gionis, R. Lempel, and Y. Maarek. When no clicks are good news. In *Industry Track talk at SIGIR'2010*, Geneva, Switzerland, 2010. See `http://www.eurospider.com/acm-sigir-industry-track-2010.html#c297`.

[9] N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC 2005 enterprise track. In *TREC*, volume 5, 2005.

[10] A. Culotta, N. K. Ravi, and J. Cutler. Predicting the demographics of twitter users from website traffic data. In *Proceedings of AAAI*. AAAI Press, 2015.

[11] D. Di Castro, L. Lewin-Eytan, Y. Maarek, R. Wolff, and E. Zohar. Enforcing k-anonymity in web mail auditing. In *Proceedings of WSDM*. ACM, 2016.

[12] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla. Inferring user demographics and social strategies in mobile social networks. In *Proceedings of KDD*. ACM, 2014.

[13] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff I've seen: A system for personal information retrieval and re-use. In *Proceedings of SIGIR*, 2003.

[14] M. Grbovic, G. Halawi, Z. Karnin, and Y. Maarek. How many folders do you really need?: Classifying email into a handful of categories. In *Proceedings of CIKM*. ACM, 2014.

[15] D. Hawking and K. Griffiths. An enterprise search paradigm based on extended query auto-completion: Do we still need search and navigation? In *Proceedings of ADCS '13*, Brisbane, Australia, 2013.

[16] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *Proceedings of WWW*. ACM, 2007.

[17] B. J. Jansen and L. Solomon. Gender demographic targeting in sponsored search. In *Proceedings of CHI*. ACM, 2010.

[18] D. Jiang, J. Pei, and H. Li. Mining search and browse logs for web search: A survey. *ACM Trans. Intell. Syst. Technol.*, 4(4), Oct. 2013.

[19] R. Jones and K. L. Klinkner. Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In *Proceedings of CIKM*. ACM, 2008.

[20] S. R. Kairam, M. R. Morris, J. Teevan, D. J. Liebling, and S. T. Dumais. Towards supporting search over trending events with social media. In *ICWSM*, 2013.

[21] E. Kharitonov and P. Serdyukov. Demographic context in web search re-ranking. In *Proceedings of CIKM*. ACM, 2012.

[22] B. Klimt and Y. Yang. The Enron corpus: A new dataset for email classification research. In *Machine learning: ECML 2004*. Springer, 2004.

[23] Q. Mei, D. Zhou, and K. Church. Query suggestion using hitting time. In *Proceedings of CIKM*. ACM, 2008.

[24] M. Shokouhi. Learning to personalize query auto-completion. In *Proceedings of SIGIR*. ACM, 2013.

[25] I. Soboroff, A. P. de Vries, and N. Craswell. Overview of the TREC 2006 enterprise track. In *TREC*, 2006.

[26] A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), Feb. 2001.

[27] J. Teevan, E. Adar, R. Jones, and M. A. S. Potts. Information re-retrieval: Repeat queries in yahoo's logs. In *Proceedings of SIGIR*. ACM, 2007.

[28] J. Teevan, D. Ramage, and M. R. Morris. #twittersearch: A comparison of microblog search and web search. In *Proceedings of WSDM*. ACM, 2011.

[29] A. Tsotsis. Comscore says you don't got mail: Web email usage declines, 59% among teens! In *Techcrunch*, Feb 2011.

[30] P. Wang, J. Guo, Y. Lan, J. Xu, and X. Cheng. Your cart tells you: Inferring demographic attributes from purchase data. In *Proceedings of WSDM*. ACM, 2016.

[31] I. Weber and C. Castillo. The demographics of Web search. In *Proceedings of SIGIR*. ACM, 2010.

[32] I. Weber and A. Jaimes. Who uses web search for what: And how. In *Proceedings of WSDM*. ACM, 2011.

[33] E. Zhong, B. Tan, K. Mo, and Q. Yang. User demographics prediction based on mobile data. *Pervasive Mob. Comput.*, 9(6), Dec. 2013.