Inferring the Demographics of Search Users

When Social Data Met Search Queries

Bin Bi*
UCLA
United States
bbi@cs.ucla.edu

Milad Shokouhi
Microsoft Research
United Kingdom
milads@microsoft.com

Michal Kosinski
University of Cambridge
United Kingdom
michal@michalkosinski.com

Thore Graepel
Microsoft Research
United Kingdom
thoreg@microsoft.com

ABSTRACT

Knowing users' views and demographic traits offers a great potential for personalizing web search results or related services such as query suggestion and query completion. Such signals however are often only available for a small fraction of search users, namely those who log in with their social network account and allow its use for personalization of search results. In this paper, we offer a solution to this problem by showing how user demographic traits such as age and gender, and even political and religious views can be efficiently and accurately inferred based on their search query histories. This is accomplished in two steps; we first train predictive models based on the publically available myPersonality dataset containing users' Facebook Likes and their demographic information. We then match Facebook Likes with search queries using Open Directory Project categories. Finally, we apply the model trained on Facebook Likes to large-scale query logs of a commercial search engine while explicitly taking into account the difference between the traits distribution in both datasets. We find that the accuracy of classifying age and gender, expressed by the area under the ROC curve (AUC), are 77% and 84% respectively for predictions based on Facebook Likes, and only degrade to 74% and 80% when based on search queries. On a US state-by-state basis we find a Pearson correlation of 0.72 for political views between the predicted scores and Gallup data, and 0.54 for affiliation with Judaism between predicted scores and data from the US Religious Landscape Survey. We conclude that it is indeed feasible to infer important demographic data of users from their query history based on labelled Likes data and believe that this approach could provide valuable information for personalization and monetization even in the absence of demographic data.

*The work was done during Bin's internship at Microsoft Research Cambridge.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]:

General Terms

Algorithms, Human Factors

Keywords

User demographics, Personalized search, Social networks

1. INTRODUCTION

In recent years, we have been witnessing the rapid emergence of social networks and an increasing amount of user generated data. Meanwhile, it became apparent that the relevance of search results can be improved by personalization, i.e., by taking into account additional information about the user, such as interests, demographic and psychological traits, social background, or the context of the search. As a consequence, search engines have been evolving into *social-aware* platforms, Google's social layer (Google+), and Bing's *social pane* being perhaps the two most noteworthy examples.

While leveraging the background information about the users in ranking models has shown significant promise in enhancing users' search experience both in academic [Carmel et al., 2009] and industrial[1] studies, obtaining such features for all users can be difficult. For instance, a recent study suggests that only about 22% of Bing users are logged into Facebook account while searching[2], and even them may have not given the search engine access to their profile information. It would therefore be useful to be able to infer characteristics of users relevant to their search experience from information more readily available in the context of a search engine, such as the search query histories.

This paper addresses the question of how demographic traits and users' views can be inferred based on the query histories. The main challenge, however, lies in the fact that only a very limited amount of data is available to allow training models for predicting such traits based on the search

[1]Google blog, `http://bit.ly/YaJvSml`
[2]Search Engine Land: `http://selnd.com/R6dpTN`

queries. For example, Microsoft user accounts provide no access to political views and religion, and only a small amount of data related to demographic traits.

How, then, can we build a machine learning system that predicts user demographics from query histories? What comes to the rescue is a substantial publicly available dataset called myPersonality[3], offering psychometric test results and contents of the Facebook profiles for millions of anonymous Facebook users who volunteered to donate their data for research purposes. In particular, myPersonality database allows matching users' demographic profiles with their Facebook Likes, i.e., those online entities and Facebook Pages[4] with which users have associated themselves using the Facebook Like button. Here we show how Facebook Likes can be used to build a model predicting users' individual traits that can be later applied to search query data.

There are two issues that need to be addressed to apply the model built on Facebook Likes to query histories. First, Facebook Likes need to be matched against queries. We achieve that by developing a common representation for Facebook Likes and search queries within the Open Directory Project (ODP)[5] categories. Second, the distribution of users' traits differs between Facebook and search samples. Traditional machine learning algorithms commonly assume that the training and test samples are randomly drawn from the same distribution. To address this issue, we design a novel learner which is able to adapt the model learned from social data to search queries with a different distribution. This learner does not require unlabelled search queries to be seen at training time, which relaxes the condition of traditional transfer learning. Experimental results show that the new learner can give high prediction accuracy as well as some interesting demographic results.

Hence, the paper makes four important contributions:

- We show how to predict users' traits based on the search query logs by applying models developed on the Facebook Likes data.
- We show how to use ODP categories to match Facebook Likes with search queries.
- We demonstrate how to mitigate the problem of differing distributions of the traits.
- We provide experimental results that show the validity of the approach by comparing the predictions with ground truth data and with aggregate data at the US state level.

The structure of the paper is as follows: we begin by discussing related work in Section 2. We then describe our ODP representation scheme, learning algorithms and approach to dataset shift mitigation in Section 3. The datasets are described in Section 4, while details of the evaluation methodology are provided in Section 5. Finally, we describe our experimental results in Sections 6 and 7, and conclude in Section 8.

[3]See `http://mypersonality.org/wiki` for more details.
[4]Facebook Pages, `http://www.facebook.com/pages`
[5]Open directory project, `http://www.dmoz.org`

2. RELATED WORK

Our work is related to a wide spectrum of previous studies ranging from inferring the demographics of individual users, to the application of user demographics in predicting global trends or individual behaviour.

The impact of demographics & personality. Bachrach et al. [2012] investigated the correlation between users' personality and the properties of their social network profiles. They showed that some personality traits such as Extroversion and Neuroticism can be accurately predicted based on the user's profile. A similar study was conducted by Quercia et al. [2011] on Twitter users.

Kosinski et al. [2012] demonstrated that there is a psychologically meaningful relationship between the users personality profiles obtained using a questionnaire, and their choice of websites extracted from Facebook Likes.

Weber and Castillo [2010] used the Yahoo! query logs and profile information to compare the queries submitted by users with different demographics. They further analysed the queries submitted from each US ZIP code separately and mapped them against the US-census information for those area codes. Their results suggested that users with similar demographics are more likely to search for similar things. Weber and Jaimes [2011] examined the queries submitted from different ZIP codes augmented by US-census data to highlight the differences in user behaviour and search patterns of various demographic groups. We take this line of previous work to the next level by showing that the demographics of users can be automatically predicted based on their past queries.

Lorigo et al. [2006] discovered that male and female users have different search behaviour; for instance, females on average submit longer queries. Jansen and Solomon [2010] found that males and females interact differently with sponsored search results.

Kharitonov and Serdyukov [2012] demonstrated how reranking the search results based on users' genders may enhance their experience in particular for ambiguous queries.

Bennett et al. [2011] inferred a compact density representation of locations of users that access different websites and showed that those features can be used for personalizing and reranking the search results.

Inferring user demographics. Torres and Weber [2011] reported that the reading levels of clicked pages are correlated with the demographic characteristics of the clicking users. Weber et al. [2012a,b] relied on user clicks on political blogs annotated with *leaning* to assign a leaning score (left versus right) to queries.

Pennacchiotti and Popescu [2011] used the linguistic content of user tweets, along with their other social features to predict the political orientation, ethnicity and the favourite business brands of Twitter users. They found the user-centric features such as linguistic content to be more effective than social graph features in their classification task.

Ying et al. [2012] showed that the users demographics can be predicted according to their mobile usage behaviour, such as the number of text messages sent or received. Otterbacher [2010] inferred the author gender of IMDB reviews based on stylistic and content features.

Jones et al. [2007] investigated the problem of inferring

users demographics based on their queries but mostly focused on the privacy angle. They leveraged bag-of-word classifiers based on queries to train their models.

Perhaps in the most similar work to ours Hu et al. [2007], predicted the users' ages and genders based on their browsing model. For each website in their corpus they used the Microsoft Live ID information of users that accessed them to build a demographic model. They then used these models to predict the ages and genders of other users that access the same website. In our approach we bring the social and query data into the same space but mapping them against the ODP categories. As a result, we have a much denser feature space that allows us to have high generalisably and cover several other interesting aspects such as religion and political views in the inference.

From query trends to global statistics. Weber and Jaimes [2010] monitored the Yahoo! query logs to determine if the same queries were submitted by different demographic groups at different times. Their analysis revealed that certain queries (e.g. movies) are searched by distinct demographics at different times, suggesting an *information flow* pattern between different groups of users.

Goel et al. [2010] used query volume to predict the opening weekend box-office revenue of films, first-month sales of video games and the ranks of songs on the Billboard Hot 100 chart. In each of these cases, the authors found that there was a significant correlation between the query volume and future outcomes. Ettredge et al. [2005] performed a similar study but focused on predicting the unemployment rate.

Ginsberg et al. [2009] accurately detected the influenza epidemics by only using the frequency and volume of certain queries in Google logs. Later on, Kong et al. [2010] utilized click-through for the same purpose, and Culotta [2010] repeated a similar analysis on Twitter data.

Domain adaptation and transfer learning. Our work is also related to domain adaptation and transfer learning techniques. In domain adaptation [Daumé and Marcu, 2006] typically the same feature space is shared by the source and target domains. We also deal with two distinct source (social data) and target (queries) spaces in our experiments, and bridge them by mapping them to a single common space (based on ODP categories).

Transfer learning techniques can be used to resolve the problem of the different distributions between source and target spaces. It is worth noting that in contrast to typical transfer learning models [Dai et al., 2007; Fan et al., 2005; Zadrozny, 2004], our approach requires neither any data sharing between the source and target domains, nor any target data to be seen at training time.

3. MODELING USER DEMOGRAPHICS

As mentioned in the Introduction, we are addressing the problem of inferring users' traits from search queries based on the models trained on an independent set of Facebook Likes and profiles. We thus face two challenges

- How can we find a common representation for search queries and Facebook Likes?
- How can we address the problem that the users' traits are distributed differently in those two datasets?

We address the first problem by mapping both search queries and Facebook Likes into a common representation given by the Open Directory categories, which form a mini-ontology of entities on the Web and can be thought of as a coarse grained representation of both search queries and Facebook Likes. Figure 1 illustrates the common representation based on the DMOZ Open Directory Project (ODP) categories. For Facebook Likes we turn the title of each *liked entity* into a query and submit it to a search engine (for example for the *lady gaga* Facebook Like, we submit the query *lady gaga*). We classify each of the top ten results returned by the search engine (Bing was used in this study) into one of the top two-levels of the DMOZ/ODP categories, assigning a maximum of three categories to each result. In total, there are 219 topical categories such as Arts/Movies, Business/Jobs and Computers/Internet. For learning the category classifiers we follow the approach described by Bennett et al. [2010] and apply logistic regression with L2 regularization on a 2008 crawl of the documents linked with the ODP index. Using the output of these classifiers we then represent each Facebook Like in the myPersonality dataset by a 219-dimensional vector. Each element of this vector denotes the number of times that a particular ODP category has been assigned to the search results returned for that Like. We then repeat the same process on search (Bing) users. To generate the topical feature vector for each user, we collect the queries from their search history and classify them in the same way as the we did for the Facebook Likes. Each user is represented again by a 219-dimensional vector, in which each element denotes the number of times the corresponding ODP category has been assigned to top-ranked documents returned for user queries. The feature values are normalized into probabilities so that they all sum to one for each user.

The second problem arises because of the differing users' traits distribution between users in the Facebook and search queries samples. Traditional machine learning algorithms commonly assume that the training data consist of samples randomly drawn from the same distribution as the test samples about which the learned model is expected to make predictions. This assumption is violated in our scenario where the model trained on Facebook data is applied to a query log to predict users' demographic characteristics in the search engine. One of the examples is that there are relatively more female user in the Facebook (myPersonality) dataset, compared to search (Bing) users. Naively training on one dataset and testing on the other can significantly decrease the predictive accuracy of a traditional learning algorithm. This is because a learning algorithm aims to learn an optimal model for the query log by minimizing the expected risk:

$$\hat{\theta} = \arg\min_{\theta} \sum_{(q,y)\in D_q} P(D_q)\ell(q, y, \theta) \quad (1)$$

where D_q is query log data, and $\ell(q, y, \theta)$ is a loss function with parameter θ.

However, since we have to assume that no labeled data is available from the query log, we have to learn a model from the Facebook data instead by minimizing the empirical risk:

$$\hat{\theta} = \arg\min_{\theta} \sum_{(l,y)\in D_f} P(D_f)\ell(l, y, \theta) \quad (2)$$

If $P(D_q) = P(D_f)$, the two optimization problems are

Figure 1: The workflow of our framework for inferring users' demographics based on the search queries. On the left, the Facebook Likes of a small group of users are mapped to their corresponding ODP categories by issuing them as queries and classifying the top search results. On the right, the search users are represented similarly by the set of ODP categories associated with the top-ranked results returned for their queries.

approximately equivalent. However, as we can observe from the comparison of the Facebook and search data (see Table 1), the two distributions $P(D_q)$ and $P(D_f)$ are different.

To predict demographic characteristics of the users in a query log, we essentially seek to obtain the conditional probability distribution $P(Y|Q, D_q)$, where Y denotes the demographic characteristic of a user who issued queries Q, and D_q denotes the query log. Note that $P(Y|Q, D_q) \neq P(Y|L, D_f)$ as discussed earlier.

Since we choose to represent each user by a probability distribution over ODP categories, $P(Y|Q, D_q)$ can be marginalized across ODP categories C:

$$P(Y|Q, D_q) = \sum_C P(Y|C, D_q)P(C|Q, D_q) \tag{3}$$

By Bayes' rule, $P(Y|C, D_q)$ is given by:

$$P(Y|C, D_q) = \frac{P(Y|D_q)P(C|Y, D_q)}{P(C|D_q)}, \tag{4}$$

where $P(Y|D_q)$ is the probability of class Y in the query log, which captures our prior knowledge about the relative frequencies of users of different demographics in a search engine. These quantities can be obtained from the search engine internal statistics, or publicly available statistics about the search users. On the other hand, $P(C|D_q)$ captures the relative frequencies of queries of category C. This quantity could be estimated from search logs, but can also be approximated from the ODP/DMOZ statistics assuming that the ODP corpus is statistically similar to the set of results returned by the search engine. $P(C|Y, D_q)$ is the probability that a user with demographics Y is interested in category C when issuing a query. The key insight here is that we can assume that whether a user is interested in some category C or not depends on their demographics Y, independent of whether he or she is using Facebook or doing search. Therefore, it is reasonable to make the conditional independence assumption that

$$P(C|Y, D_q) = P(C|Y) = P(C|Y, D_f)\,, \tag{5}$$

which means that $P(C|Y, D_q)$ can be estimated from Facebook data D_f. Let θ^Y denote the probability distribution $P(C|Y)$. In order to avoid problems of estimation due to sparsity of the data we estimate the parameter vector θ^Y using Bayesian Maximum A Posteriori (MAP) estimation. In particular, we estimate θ^Y by:

$$\hat{\theta}^Y = \arg\max_{\theta^Y} P(\theta^Y|D_f) = \arg\max_{\theta^Y} P(D_f|\theta^Y)P(\theta^Y)\,. \tag{6}$$

This is a standard Bayesian estimation problem with a multinomial likelihood and a conjugate Dirichlet prior $P(\theta^Y)$ parameterized by pseudo-counts $\{\alpha_k\}$, $(\alpha = \sum_k \alpha_k)$. If there is prior knowledge available, this can be taken into account, otherwise one can initialize the pseudo counts $\{\alpha_k\}$ uniformly. The resulting MAP solution is given by:

$$\theta_k^Y = \frac{N_k^Y + \alpha_k - 1}{N^Y + \alpha - K}, \tag{7}$$

where N_k^Y is the number of times the webpages, which are returned for the Likes of users of class Y, fall into the kth category, K is the total number of ODP categories, and N^Y is the total number of categories for webpages returned for the Likes of users of class Y. Note that we estimate the probability $P(C|Q, D_q)$ in Equation (3) in a similar way.

In summary, the methodology outlined above allows us to train a demographics classifier on users characterized by their collection of Facebook Likes, yet evaluate it on users characterized by their search query history. We believe that the two key ideas of a) creating a common representation in terms of ODP, and b) of mitigating the data shift problem by breaking up the problem into separate estimation tasks for demographics given category and category given query history will be more generally applicable to problems in which labels are available, but are not directly linked with the representation of interest through suitable training data.

4. DATA

myPersonality Dataset (Facebook). The myPersonality dataset was collected through the *myPersonality* Facebook

Table 1: The distribution of age and gender in search queries and Facebook Likes datasets.

Dataset	Teenage (10-18)	Youngster (19-24)	Young (25-34)	Mid-Age (35-49)	Elder (50+)	Male	Female
Social Dataset	3%	49%	32%	14%	2%	37%	63%
Search Dataset	2%	11%	24%	39%	24%	53%	47%

application, which allowed its users to take real psychometric tests and receive feedback on their scores. In addition to the results of the tests, respondents could opt in to record their Facebook profile data to be used for the research purposes. myPersonality contains detailed psycho-demographic profiles of more than 6 million unique users from diverse age groups, backgrounds, and cultures. Respondents were motivated to answer honestly, as the only gratification they received for their participation was feedback on their results. We used a subset of myPersonality users from US described by their age, gender, political views, religion, and lists of their Facebook Likes. We filter out all Facebook Likes associated with less than ten users. The resulting dataset contains over 457,000 users, 122,000 unique Likes, and over 11 million associations between the users and Facebook Likes. Users' religion and political views were stored as free text. Although the great majority of users simply have the typical religion/party/philosophy names in those fields (e.g. Christian, Liberal), sometimes we had to use regular expression matching to extract the relevant information. For instance, "Christian - Baptist" was recoded as "Christian" and "I dont go to church because i wanna leave room in the pews for the sinners that need it -mr. magee" was ignored after mismatching all of our regular expressions.

Bing Query Logs (Search). We apply the models trained on the myPersonality dataset to infer the traits of users characterized by search queries. Search query logs were obtained from Bing and were collected between October 14, 2012 and October 28, 2012. We have selected queries submitted by the US users that were signed in with their Microsoft Live account while issuing their queries. In total, we have collected 133 million queries from 3.3 million unique users. Each user was also described by age and gender as reported in their Microsoft Live profiles.[6]

Differing distributions (Data Shift). Table 1 shows that the distributions of user demographics significantly differ between myPersonality and search query logs datasets. For instance, on average there are more young and female users in our Facebook data, which considering the nature of myPersonality test may not be surprising.[7]

[6]In both samples only anonymous data was used. The user IDs were all annoymized such that the actual usernames could not be identified.

[7]It is important to note that the demographics reported here for our search and social datasets necessarily cannot be regarded as representative statistics for Bing and Facebook. The distributions in the datasets, particularly for the myPersonality data, are significantly affected by how the data is collected. The unique characteristics of the myPersonality test is likely to attract certain types of audience more than others. Readers are encouraged to refer to other sources (such as `alexa.com`) for more representative statistics.

5. EVALUATION

For each user trait used here, we first train a model on 66% of Facebook users in myPersonality dataset and test it on the remaining 34%. We then apply the same model on search queries and repeat the classification for search users.

Evaluation on myPersoanality Sample. In the binary classification tasks such as predicting gender or political view (liberal vs. conservative) we use the area under the ROC curve (AUC) measure for evaluating accuracy. The ROC curves are created by plotting the ratio of true positive rate versus false positive rate at various threshold settings. We turn each of the multiclass classification tasks such as predicting religion (among Christian, Buddhist, Jew, Agnostic) into multiple binary classification problems (e.g. Buddhist or not Buddhist) and report the average values at the end.

Evaluation on Bing sample. The age and gender information of the Bing users was obtained from their Microsoft Live profiles. Hence, we can repeat the same type of AUC evaluation, but this time with the labels coming from Microsoft Live accounts.

Religion and political views are not available in the Microsoft Live profiles, hence we do no have the ground-truth information on the individual user level. Therefore, we evaluate the accuracy of the trained classifiers on how well their output matches the officially reported state-level statistics. We first classify the religion and political views of individual users (e.g. religion = Christian) and aggregate those results on the state level (e.g. 74% Christians in California) by using users' location acquired from the IP address. We then look up the corresponding reported values for each state from publicly available official statistics (e.g., what percentage of Californians are Christian). Next, for each given class (e.g., Christianity) and each state, we calculate the percentage of search users that are classified in that category with respect to (1) our predictions and (2) official statistics. Finally, we compute the Pearson correlation value (ρ) between (1) and (2) and consider it as a proxy for the accuracy of the prediction.

6. EXPERIMENTS

Using the compact ODP representation described earlier, we managed to model all users in both Facebook and search queries datasets. In comparison, an exact-match approach that compares the text of queries and Likes finds only 5.3% overlap by which only 36% of search users can be modelled and even for those there are often only few non-zero features.

Table 2 displays the evaluation results of the classifiers built on Facebook sample for inferring different demographics. The middle column (Facebook-Facebook) shows the AUC values when we trained and tested on the Facebook dataset. The right column (Facebook-Search) shows the Facebook model accuracy on classifying Search users.

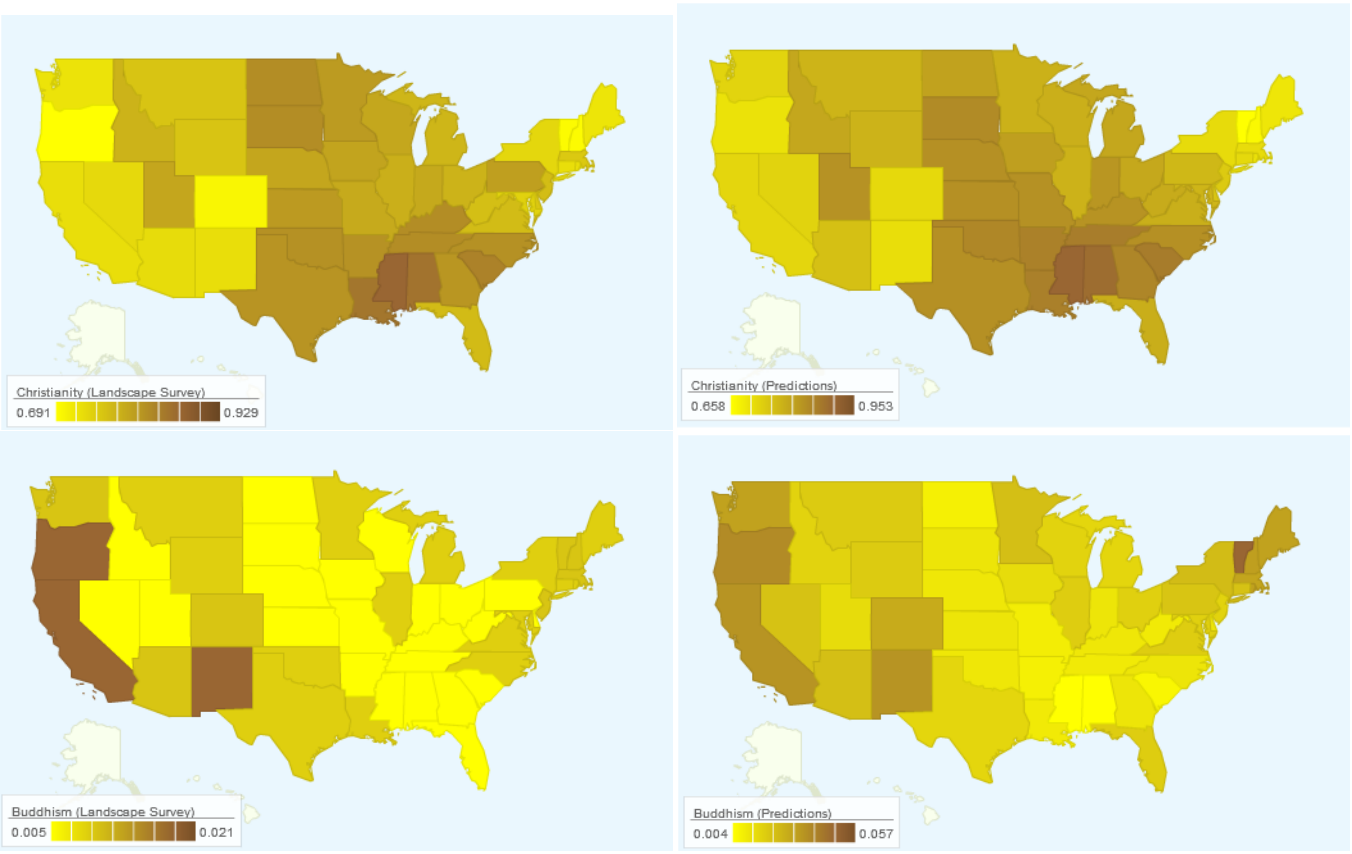

Figure 2: (Top-Left) The distribution of Christians in the *Contiguous* United States according to the U.S. Religious Landscape Survey. (Top-Right) The distribution of Christians in the US as predicted based on user queries. The Pearson correlation (ρ) is 0.39. (Bottom-Left) The distribution of Buddhism in the *Contiguous* United States according to the U.S. Religious Landscape Survey. (Bottom-Right) The distribution of Buddhism in the US as predicted based on user queries. The Pearson correlation (ρ) is 0.53. The spectrum bar at the left corner of each map specifies the scale and the corresponding color codes.

Table 2: The Area under the ROC Curve (AUC) for different demographic prediction models. The numbers in the middle column show the accuracy of a model trained on Facebook data for predicting the demographics of Facebook users. In the right column, the models trained based on Facebook data are tested on search query sample. The missing values "-" are used where the per-user ground-truth information is not available for AUC evaluation.

AUC	Facebook-Facebook	Facebook-Search
Gender	0.836	0.803
Age	0.771	0.735
Religion	0.758	-
Political view	0.739	-

For gender classification, we train two separate classifiers; one for male, and one for female, each computing the probability of given gender based on the user profile (ODP features). Each user is compared against both of these classifiers, and the one producing the highest probability is used to set the class of gender. The results in Table 2 show that the classification reaches 83% and 80% accuracy respectively when tested on Facebook and Search samples. Not surprisingly, the accuracy of a model trained based on the Facebook sample, is higher when it is tested on other users from the same dataset. However, the relative loss is not substantial, particularly considering the significant differences in the demographic distributions of these two sets (37% Male in Facebook dataset, compared to 53% among Search users).

For age classification, we grouped the users in each dataset into five separate age groups as listed in Table 1. For each age group we compute a model based on the training subset of users in our Facebook dataset. At testing, each user – in Facebook and Search datasets – is compared against these models, and the one producing the highest probability is used for classification. On the testing subset of Facebook users, the trained classifier achieves 77% accuracy, while this number is slightly lower (73.5%) when applying the model on the Search sample.

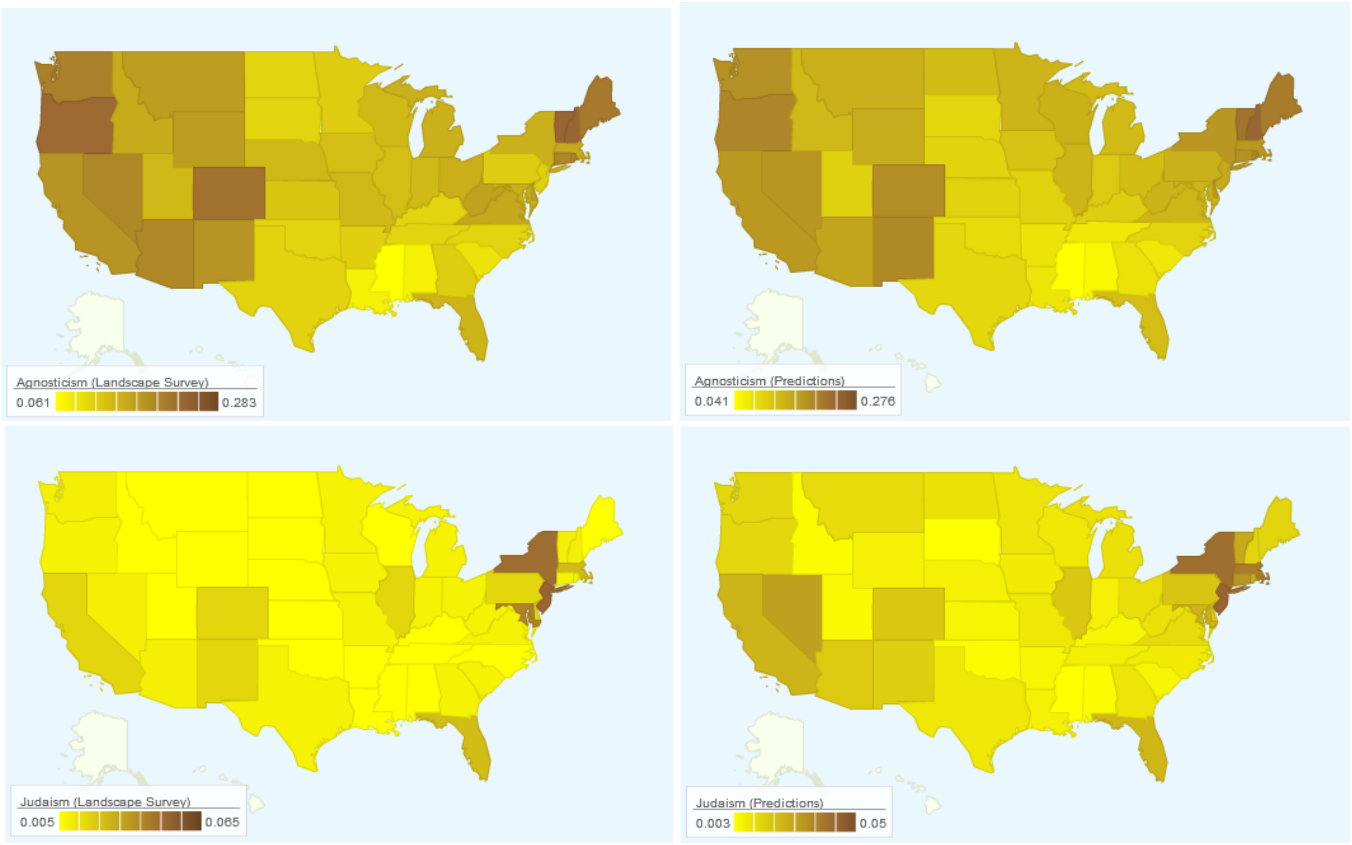

Figure 3: (Top-Left) The distribution of Agnostics in the *Contiguous* United States according to the U.S. Religious Landscape Survey. (Top-Right) The distribution of Agnostics in the US as predicted based on user queries. The Pearson correlation (ρ) is 0.27. (Bottom-Left) The distribution of Judaism in the *Contiguous* United States according to the U.S. Religious Landscape Survey. (Bottom-Right) The distribution of Judaism in the US as predicted based on user queries. The Pearson correlation (ρ) is 0.54. The spectrum bar at the left corner of each map specifies the scale and the corresponding color codes.

To classify users' religion, we first apply a set of regular expressions as described in Section 4 to assign the social users into four groups: {Christian, Jewish, Buddhist, and Agnostic/Unaffiliated }. These are also the four major religions in the United States according to U.S. Religious Landscape Survey,[8] accounting respectivelly for of 78.4%, 1.7%, 0.7%, and 16.1% of the entire US population. We use these nationwide statistics as the prior when classifying the users in the Search dataset. The AUC while classifying users religion in the testing subset of Facebook dataset is 76%. Importantly, as there is no information about the religion on the Search user level, thus, the accuracy of the classification was evaluated in terms of how well it predicted the state-level distributions as described in Section 4.

Figure 2 depicts the state distributions of Christians (top) and Buddhists (bottom) according to the U.S. Religious Landscape Survey on the left, and according to our predictions for the search engine users on the right.[9] The models predict that depending on the state, 65.8%-95.3% of search users are Christians. These values are comparable to 69.1%-92.9% reported in the Landscape Survey. We also correctly identify the Mississippi state as the one with the highest ratio of Christians, and the states on the east coast with the lowest density ($\rho = 0.94$). Similarly the models predict 0.4%-5.7% of search users in the dataset to be Buddhist depending on the state, which is not far from the 0.5%-2.1% range reported in the Landscape Survey. The models predict Vermont, Oregon, California, and New Mexico to have the largest population of Buddhists, and apart from the former – that accounts for 0.001% of our dataset and hence is somewhat prone to noise – the remaining three are also listed as the top three Buddhist states in the Landscape Survey (Overall, $\rho = 0.72$).

Figure 3 demonstrates the spread of Agnostic (top) and Jewish (Bottom) people in the United States. The models predict 4.1%%-27.6% of the search users in our dataset to be agnostic or unaffiliated with any particular religion. The official numbers from the Landscape Survey for this category lie closely between 6.1% and 28.3%. Consistent with the Landscape Survey, our models predict higher density of agnostics in North East and West, with the state of New Hampshire appearing on top of both – survey and predicted – lists ($\rho = 0.91$). According to our predictions based on search engine users, Jews account for 0.3%-5.0% of the US

[8] `http://religions.pewforum.org`

[9] The states of Alaska and Hawaii do not appear on the Landscape Survey and hence are dropped from the analysis.

population depending on the state. These numbers are fairly consistent with the 0.5%-6.5% reported on the Landscape Survey ($\rho = 0.79$). We also correctly identify the states in the North East, in particular New York to have the highest density of Jewish people. This is yet again aligned with the Landscape Survey and historical documents about the Jewish settlements in the United States.[10]

We matched a set of regular expressions against the *Political view* field of users in the Facebook dataset to group them into *liberal* (34%) and *conservative* (66%) categories. We ignored users that did not match any of the regular expressions in building our models. The distribution of liberal versus conservative in the social dataset is remarkably close to those reported by independent sources such as Gallup survey which reported 20.6% liberals versus 40% conservatives nationwide – the remainder of people in the poll were assigned to *moderate* and other groups.[11]

As in previous experiments, we build the classifiers based on the ODP features of the users in the training subset (64%) of Facebook dataset. Applying the model on the remaining (34%) of users in that dataset produces the AUC of 0.74. We then apply the same model on the Search sample; the middle and bottom maps in Figure 6 illustrate the distribution of liberals and conservatives in the US. The middle map is generated based on the per-state statistics reported by the Gallup survey. The bottom map is generated by applying the classifier trained on the Facebook sample to Search users. The predicted class for each individual user contributes to generate the overall distribution for each of the states.

To enhance the visualization, the plots were produced with respect to the nationwide average so that the differences between states become more prominent. For instance, -0.10 would mean 10% more liberal, while 0.05 would suggest 5% more conservative than the nationwide average. The middle and bottom maps in Figure 6 reveal very similar distributions (ρ= 0.72). As expected, both maps look more blue on the East-West coasts, and more red in the so-called *Bible Belt* states. Oregon with an officially reported 13.8% swing towards liberals is the most noticeable mispredicted state; this was affected to some extend by the ambiguity of the queries related to *the civil war*, a college football rivalry in Oregon, which was particularly trendy during our sampling period.

It is commonly known that liberals are more likely to vote for the Democratic Party and conservatives are more likely to vote for Republicans.[12] Thus, perhaps it is not entirely surprising to find similarities in how the states were split between Democrats and Republicans in the recent 2012 US presidential election (top map in Figure 6).

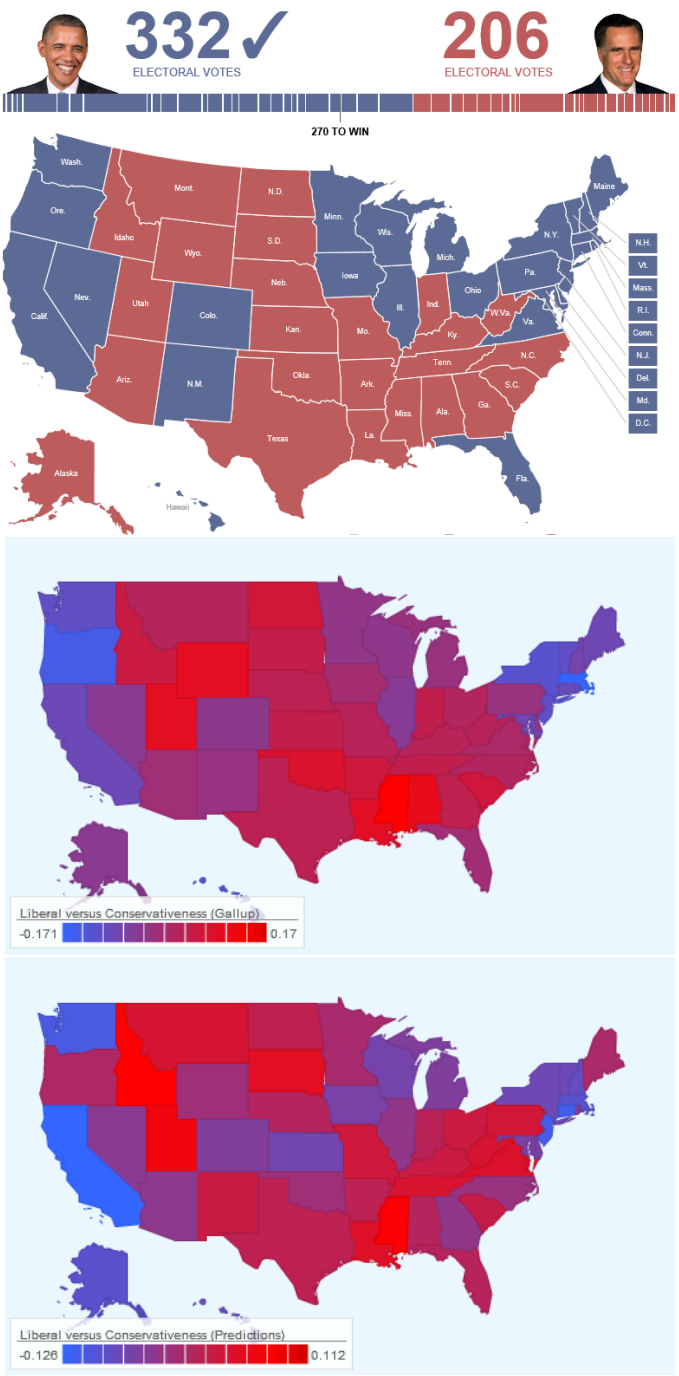

Figure 4: (Top) The outcome of 2012 the US presidential election according to The Huffington Post. The blue states were won by Democrats and the red states by Republicans. (Middle) The distribution of conservatives versus liberals according to an independent poll – Gallup.[a] (Bottom) Liberal vs. conservative predictions on Bing users based on the models learned according to Facebook data. The Pearson correlation (ρ) between the Gallup data and our per-state predictions is 0.72. The spectrum bar at the left corner of each map specifies the scale and the corresponding color codes.

7. IMPORTANCE OF ODP CATEGORIES

In this section we show the importance of each category in predicting a given type according to its *information gain* computed in a leave-one-out fashion. That is, for each ODP category C (e.g. Arts/Movies), and a given demographic type Y (e.g. Gender), we first calculate the prior values according to all other 218 categories in our data, and then calculate the change in information entropy when C is con-

[10] `http://en.wikipedia.org/wiki/American_Jews`
[11] Gallup poll, `http://bit.ly/hsceKj`
[12] Gallup Politics, `http://bit.ly/AoyIg4`, and Rasmussen Report, `http://bit.ly/L85SmV`

[a] Source: Gallup, `http://bit.ly/zCOPHk`

Table 3: The ODP categories with the highest information gain for different types of demographics.

Gender	Age	Religion	Political view
Sports/Basketball	Arts/Movies	Religion_and_Spirituality/Christianity	Politics/Liberalism
Games	Computers/Data_Communications	Religion_and_Spirituality/Religious_Studies	Politics/Conservatism
Sports/Soccer	Games	Religion_and_Spirituality/Scientology	Society/History
Shopping/Gifts	Shopping/Toys_and_Games	Society/History	Arts/Movies
Shopping/Jewellery	Computers/Software	News/Media	Science/Social_Sciences

sidered as,

$$IG(Y, C) = H(Y) - H(Y|C) \tag{8}$$

Here, $H(Y)$ represents the prior entropy for the demographic type Y across all users, and $H(Y|C)$ is the same value conditioned on observing category C in the user's profile. Table 3 shows the categories with the highest information gains for classifying each of the demographics. For classifying gender, sport and shopping related categories are most effective. Art/Movies, Games, Shopping/Toys_and_Games and computer-related categories are best in discriminating between different age groups. For religion, subcategories of Religion_and_Spirtuality are the most important features, and for politics – not surprisingly – Politics/Liberalism and Politics/Conservatism have the highest information gain values for distinguishing between liberals and conservatives.

We also calculate the *influence* (β) of each category $c \in C$ (e.g. Arts/Movies) in classifying a given demographic type to a particular class $y \in Y$ (e.g. Gender = Male) by,

$$\beta = \frac{P(C = c|Y = y) - \mu}{\mu} \tag{9}$$

where μ represents the average probability of class $c \in C$, for all values of $y \in Y$. That is,

$$\mu = \frac{\sum_{y \in Y} P(c|y)}{|Y|} \tag{10}$$

When ranking categories by Equation (9), for gender, we found Shopping/{Jewelry, Health, Pets, Craft}, Arts/Design, and Society/Relationship as the most *influential* categories for classifying females. For males, Shopping/Gift, Sports sub-categories, Games and Recreation/Guns had the highest influence. For political views, Politics/Conservatism, and Society/{Military, Politics, Religion_and_Spirituality} had the highest β scores for conservatives, while for liberals Society/Gay,_Lesbian,_and_Bisexual, Politics/Liberalism, and Computers/Artificial_Intelligence were ranked highest.

For age, Kids_and_Teens/Health had the highest β among teenagers. Adult/Society and Sports/Wrestling were the highest-ranked categories for youngs and youngsters. Shopping/Jewellery, and Business/Hospitality were closely ranked on top for mid-age users, while Shopping/Ethnic_and_Regional and News/Media were the top two for elders.

Finally for religion, Religion_and_Spirituality/Christianity, and Religion_and_Spirituality/Scientology had respectively the highest bias towards Christians and Buddhists. For Jews, somewhat surprisingly Computers/Computer_Science was ranked highest, while agnostics had the strongest negative biases towards Religion_and_Spirituality/Religious_Studies, and Religion_and_Spirituality/Scientology.

8. CONCLUSIONS

In this paper, we addressed the problem of inferring users traits – namely age, gender, religion and political view – from their search queries. We trained our predictive models on a sample of Facebook users that had agreed to provide their Likes and other profile information for research purposes. To the best of our knowledge, this is the first study that infers the demographics of search users based on the models trained on the independent social datasets.

We demonstrated that both Facebook Likes and search queries can be translated into a common representation via mapping to ODP categories. In addition, we addressed the data-shift problem by breaking up the problem into separate estimation tasks for demographics given category, and category given query history.

Our experimental results on a large scale query log of a commercial search engine confirms that the demographics of search users can be accurately predicted based on models trained on an independent social data. The trained classifiers achieved 80% and 74% AUC respectively for classifying gender and age. For various religious and political views the models consistently ranked the US states close to their rankings reported in the official statistics (Pearson $\rho > 0.72$ in all our experiments).

For future work, we are interested in expanding the models to capture other types of user traits, such as personality, intelligence, happiness, or interests and measuring the applications of those inferred traits in personalization, reranking and monetization of the search results.

References

Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell. Personality and patterns of Facebook usage. In *Proceedings of the 3rd Annual ACM Web Science Conference*, WebSci '12, pages 24–32, Evanston, IL, 2012. ACM. ISBN 978-1-4503-1228-8.

P. N. Bennett, K. Svore, and S. T. Dumais. Classification-enhanced ranking. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 111–120, Raleigh, NC, 2010. ACM. ISBN 978-1-60558-799-8.

P. N. Bennett, F. Radlinski, R. W. White, and E. Yilmaz. Inferring and using location metadata to personalize web search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 135–144, Beijing, China, 2011. ACM. ISBN 978-1-4503-0757-4.

D. Carmel, N. Zwerdling, I. Guy, S. Ofek-Koifman, N. Har'el, I. Ronen, E. Uziel, S. Yogev, and S. Chernov. Personalized social search based on the user's social network. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages

1227–1236, Hong Kong, China, 2009. ACM. ISBN 978-1-60558-512-3.

A. Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, Washington, DC, 2010. ACM.

W. Dai, G.-R. Xue, Q. Yang, and Y. Yu. Transferring naive Bayes classifiers for text classification. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 1*, AAAI'07, pages 540–545, Vancouver, BC, 2007. AAAI Press. ISBN 978-1-57735-323-2.

H. Daumé, III and D. Marcu. Domain adaptation for statistical classifiers. *J. Artif. Int. Res.*, 26(1):101–126, May 2006. ISSN 1076-9757.

M. Ettredge, J. Gerdes, and G. Karuga. Using web-based search data to predict macroeconomic statistics. *Commun. ACM*, 48(11):87–92, Nov. 2005. ISSN 0001-0782.

W. Fan, I. Davidson, B. Zadrozny, and P. S. Yu. An improved categorization of classifier's sensitivity on sample selection bias. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, ICDM '05, pages 605–608, Washington, DC, USA, 2005.

J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, Feb. 2009. ISSN 1476-4687.

S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts. Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences*, 107(41): 17486–17490, Oct. 2010.

J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 151–160, Banff, AB, 2007. ACM.

B. J. Jansen and L. Solomon. Gender demographic targeting in sponsored search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 831–840, Atlanta, GA, 2010. ACM.

R. Jones, R. Kumar, B. Pang, and A. Tomkins. "I know what you did last summer": query logs and user privacy. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 909–914, Lisbon, Portugal, 2007. ACM.

E. Kharitonov and P. Serdyukov. Gender-aware re-ranking. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 1081–1082, Portland, OR, 2012. ACM. ISBN 978-1-4503-1472-5.

W. Kong, Y. Liu, S. Ma, and L. Ru. Detecting epidemic tendency by mining search logs. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1133–1134, Raleigh, NC, 2010. ACM.

M. Kosinski, P. Kohli, D. Stillwell, Y. Bachrach, and T. Graepel. Personality and website choice. In *Proceedings of the 3rd Annual ACM Web Science Conference*, WebSci '12, Evanston, IL, 2012.

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. *Inf. Process. Manage.*, 42(4):1123–1131, July 2006. ISSN 0306-4573.

J. Otterbacher. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 369–378, Toronto, ON, 2010. ACM. ISBN 978-1-4503-0099-5.

M. Pennacchiotti and A.-M. Popescu. Democrats, Republicans and Starbucks afficionados: user classification in Twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 430–438, San Diego, CA, 2011. ACM. ISBN 978-1-4503-0813-7.

D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. Our Twitter profiles, our selves: Predicting personality with Twitter. In *PASSAT/SocialCom 2011*, pages 180–185, Boston, MA, 2011. IEEE. ISBN 978-1-4577-1931-8.

S. Torres and I. Weber. What and how children search on the web. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 393–402, Glasgow, UK, 2011. ACM. ISBN 978-1-4503-0717-8.

I. Weber and C. Castillo. The demographics of web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 523–530, Geneva, Switzerland, 2010. ACM. ISBN 978-1-4503-0153-4.

I. Weber and A. Jaimes. Demographic information flows. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1521–1524, Toronto, ON, 2010. ACM. ISBN 978-1-4503-0099-5.

I. Weber and A. Jaimes. Who uses web search for what: and how. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 15–24, Hong Kong, China, 2011. ACM. ISBN 978-1-4503-0493-1.

I. Weber, V. R. K. Garimella, and E. Borra. Mining web query logs to analyze political issues. In *Proceedings of the 3rd Annual ACM Web Science Conference*, WebSci '12, pages 330–334, Evanston, IL, 2012a. ACM. ISBN 978-1-4503-1228-8.

I. Weber, V. R. K. Garimella, and E. Borra. Political search trends. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 1012–1012, Portland, OR, 2012b. ACM. ISBN 978-1-4503-1472-5.

J. J.-C. Ying, Y.-J. Chang, C.-M. Huang, and V. S. Tseng. Demographic prediction based on users mobile behaviors. In *Mobile Data Challenge 2012 (by Nokia) Workshop*, Newcastle, UK., 2012.

B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 114–, Banff, AB, 2004. ACM. ISBN 1-58113-838-5.