Neural Demographic Prediction using Search Query

Chuhan Wu
Electronic Engineering
Tsinghua University
wuch15@mails.tsinghua.edu.cn

Fangzhao Wu
Microsoft Research Asia
Beijing, China
wufangzhao@gmail.com

Junxin Liu
Electronic Engineering
Tsinghua University
ljx16@mails.tsinghua.edu.cn

Shaojian He
Microsoft Inc.
Seattle, United States
shaojh@microsoft.com

Yongfeng Huang
Electronic Engineering
Tsinghua University
yfhuang@tsinghua.edu.cn

Xing Xie
Microsoft Research Asia
Beijing, China
xingx@microsoft.com

ABSTRACT

Demographics of online users such as age and gender play an important role in personalized web applications. However, it is difficult to directly obtain the demographic information of online users. Luckily, search queries can cover many online users and the search queries from users with different demographics usually have some difference in contents and writing styles. Thus, search queries can provide useful clues for demographic prediction. In this paper, we study predicting users' demographics based on their search queries, and propose a neural approach for this task. Since search queries can be very noisy and many of them are not useful, instead of combining all queries together for user representation, in our approach we propose a hierarchical user representation with attention (HURA) model to learn informative user representations from their search queries. Our HURA model first learns representations for search queries from words using a word encoder, which consists of a CNN network and a word-level attention network to select important words. Then we learn representations of users based on the representations of their search queries using a query encoder, which contains a CNN network to capture the local contexts of search queries and a query-level attention network to select informative search queries for demographic prediction. Experiments on two real-world datasets validate that our approach can effectively improve the performance of search query based age and gender prediction and consistently outperform many baseline methods.

KEYWORDS

User Modeling, Demographic Prediction, Search Query

ACM Reference Format:
Chuhan Wu, Fangzhao Wu, Junxin Liu, Shaojian He, Yongfeng Huang, and Xing Xie. 2019. Neural Demographic Prediction using Search Query. In *The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), February 11–15, 2019, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3289600.3291034

WSDM '19, February 11–15, 2019, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-5940-5/19/02…$15.00
https://doi.org/10.1145/3289600.3291034

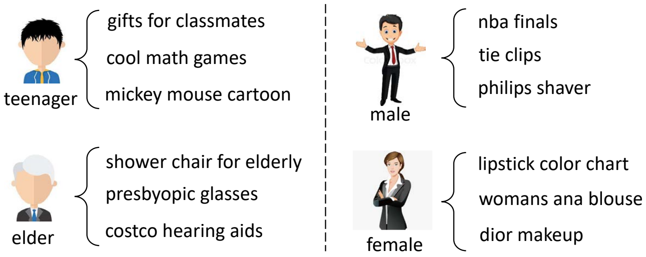
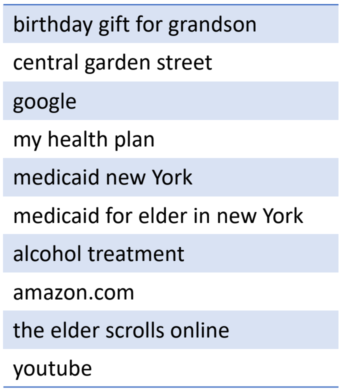


Figure 1: Search queries from users of different demographics.

1 INTRODUCTION

The demographic information of online users such as age and gender plays an important role in many personalized web applications, such as personalized search engine, recommendation and advertising [3, 9, 12, 38, 40]. For example, with the gender information of online users, the advertisers can display dress Ads to female users and shaver Ads to male users. Without the demographic information of users advertisers may show retirement insurance Ads to a teenager user and lipstick Ads to male user, which may be not effective. However, it is very difficult to obtain the user demographic information and the demographics of many online users are not available, which limits the application of demographic information in personalized web services [12, 30]. Luckily, online users frequently use search engines to search for desired information, and the search queries accumulated by commercial search engines can cover a huge number of users. In addition, the search queries from users with different demographics usually have some difference in their contents and writing styles. For example, as shown in Fig. 1 teenagers may search for "math games" and "gifts for classmates", while elders may search for "hearing aids" and "presbyopic glasses". Male users may search for "philips shaver" and female users may search for "lipstick color chart". In addition, young users are more likely to use words like "cool" in their search queries than elder users. Therefore, the search queries generated by online users can provide useful clues for inferring their demographics. Thus, in this paper we study the prediction of user demographics based on search queries.

Predicting demographics from user generated texts have been studied for several years in both data mining and natural language processing fields [19, 24, 27, 28, 33]. For example, Nguyen et al. [26]



birthday gift for grandson
central garden street
google
my health plan
medicaid new York
medicaid for elder in new York
alcohol treatment
amazon.com
the elder scrolls online
youtube

Figure 2: A motivating example for our approach. These queries are from the same user and sorted by timestamps.

proposed a linear regression model with Lasso regularization to predict users' ages from blogs, forum posts and transcribed telephone speeches. Rosenthal and McKeown [32] proposed to use logistic regression to predict user ages from blogs. Besides the blog content, they also incorporated stylistic features and behavior features in their method. Fink et al. [8] used support vector machine to predict the genders of Twitter users from their tweets. Wang et al. [36] applied long short-term memory network (LSTM) [11] to jointly predict the genders, ages and professions of social media users from their microblogging messages. However, exiting methods for user demographic prediction mainly focus on blogs, social media messages and forum posts, and the research on search query based demographic prediction is quite limited. In addition, the existing methods usually combine all the texts generated by the same user together to build user representation vector, and cannot distinguish informative texts from noisy texts for demographic prediction.

Our approach is motivated by following observations. First, not all search queries are useful for demographic prediction, and many queries are irrelevant and even noisy. For example, in Fig. 2 the queries “amazon.com” and “youtube” are not informative for age prediction. Thus, it is important to distinguish important queries (e.g., “birthday gift for grandson” for age prediction) from noisy queries (e.g., “amazon.com”) to build informative user representations for demographic prediction. Second, neighbouring search queries may have relatedness with each other [5, 22, 34], e.g., “medicaid new York” and “medicaid for elder in new York ” in Fig. 2. It is probably because they are different formulations of the same search intent. However, the relatedness between search queries with long time span is usually very weak. Modeling the local contexts of search queries is useful for learning more accurate and robust representations of queries, since search queries are usually very short and the context information in each single query is limited. Third, different words in the same query may have different importance for demographic prediction. For example, in the search query “birthday gift for grandson” the word “grandson” is more useful than “gift” for age prediction. In addition, the same word in different queries may also have different usefulness. For instance, the word “elder” is important in the search query “medicaid for elder in new York ” but is less informative in the query “the elder scrolls online”. Thus, it is important to recognize and highlight important words according to contexts to learn more informative representations of search queries.

In this paper, we propose a neural hierarchical user representation with attention (HURA) approach for demographic prediction based on search queries. Since different search queries have different informativeness for demographic prediction, instead of combining all queries from the same user into a long text for user representation, in HURA we use a hierarchical neural model to learn more informative user representations for demographic prediction. Our HURA model first learns a representation vector for each search query from words using a word encoder, which consists of a CNN network and a word-level attention network to select important words. Then we learn a representation vector for each user based on the representations of his/her search queries using a query encoder, which contains a CNN network to capture the local contexts of queries and a query-level attention network to select informative search queries for demographic prediction. Extensive experiments are conducted on two real-world search query datasets for age and gender prediction. Experimental results show that our approach can effectively improve the performance of search query based age and gender prediction and consistently outperform many baseline methods.

The rest of this paper is organized as follows. In Section 2, we briefly introduce several related works on demographic prediction. In Section 3, we introduce our HURA approach. In Section 4, we report the experimental results. In Section 5 we conclude this paper.

2 RELATED WORK

User demographic prediction has attracted many attentions from both data mining and natural language processing fields [10, 18, 29, 31]. Existing methods for demographic prediction are usually based on blogs, forum posts, social media messages, and online behaviors [7, 12, 16, 20, 21, 23, 25–28, 32, 41], and machine learning techniques are widely used in these methods. For example, Rosenthal and McKeown [32] used logistic regression to predict user ages from blogs. They extracted various features to represent users, such as blog content features (e.g., words and POS collocations), stylistic features (e.g., slang), behavior features (e.g., number of comments) and interests. Nguyen et al. [26] proposed a linear regression model with Lasso regularization to predict user ages using different kinds of texts, such as blogs, forum posts and transcribed telephone speeches. Hu et al. [12] proposed to predict users' ages and genders using their browsing behaviors. They built a bipartite graph between users and webpages using click-though data. However, the textual content of webpages is not exploited in their method. Nguyen et al. [25] applied logistic regression to predicting the ages of Twitter users according to their tweets. Peersman et al. [28] proposed to use SVM for gender prediction of Twitter users based on the content of their tweets. In recent years several deep learning based methods have been proposed for demographic prediction. For example, Zhang et al. [41] used LSTM to predict the genders and ages of social media users based on their microblogging messages. Besides the original messages, they also incorporated the retweeted messages, the comments from others and the comments to others, to learn unified user representations. Wang et al. [37] proposed a

CNN based method to simultaneously predict the ages and genders of social media users using their messages. Farnadi et al. [6] proposed a multimodal fusion model to incorporate the texts, images and user relations to predict the ages and genders of Facebook users. Different from existing methods which use blogs, forum posts and social media messages to predict demographics, our approach is based on search queries. On one hand, search queries are very common among online users and may be able to cover more online users. On the other hand, compared with other texts such as blogs and forum posts, search queries are usually very short and contain only a few keywords. Thus, the context information in search queries is usually limited, making it challenging for demographic prediction. In addition, existing methods for demographic prediction usually combine the texts from the same user together for user representation, and they cannot distinguish informative texts from noisy texts. Since the content in search queries is extremely various and many queries are irrelevant and even noisy for demographic prediction, these existing methods may be suboptimal for search query based demographic prediction.

There are only a few researches on search query based demographic prediction [1]. Bi et al. [1] proposed a method to predict the demographics of search engine users based on their search queries which combines both social network data and search query data. They firstly matched Facebook Likes data with search queries using Open Directory Project (ODP) categories. Then they trained the demographic prediction model on the labeled Facebook Likes data and applied it to predict user demographics on search queries. However, not all search queries can be mapped to ODP categories. Thus, many useful queries cannot be incorporated in their method. In addition, their method relies on the external Facebook Likes data for demographic prediction, and the coverage of users may be limited. Different from [1], our approach directly trains demographic prediction model from raw search queries. Thus, our approach can exploit all available search queries. In addition, our approach does not rely on any external social network data.

3 OUR APPROACH

In this section we present our neural hierarchical user representation with attention (HURA) approach for user demographic prediction in detail. The goal of our approach is to classify a user u into a demographic category y (e.g., an age group in age prediction and a gender category in gender prediction) based on the queries generated by this user in a certain period of time, i.e., $[q_1, q_2, ..., q_N]$, where N is the number of search queries and these queries are sorted by timestamps. Each search query is a short text, which usually consists of a few words, such as "toyota corolla 2018" and "car washes".

Our HURA approach can be decomposed into two modules, i.e., *user representation* and *user classification.* The *user representation* module aims to learn a hidden representation for each user based on their search queries, and the *user classification* module aims to classify each user into different demographic categories according to their hidden representations. Since different search queries may have different informativeness for inferring user demographic, instead of simply concatenating all search queries together, in our HURA approach we use a hierarchical representation model to learn user representations from search queries. It contains two major modules, i.e., a *word encoder* to learn query representations from words, and a *query encoder* to learn user representations from queries. The overall framework of our HURA approach is illustrated in Fig. 3.

3.1 Word Encoder

The *word encoder* module in our HURA approach is used to learn a hidden representation for a query q from its words $[w_1, w_2, ..., w_M]$. It contains three major layers.

The first layer is word embedding. Through this layer each word w_i is mapped to a low-dimensional dense vector $\mathbf{w}_i \in \mathcal{R}^D$ using an word embedding lookup table $\mathbf{E} \in \mathcal{R}^{V \times D}$, where V is vocabulary size and D is the embedding dimension. The parameters of this word embedding lookup table are tuned during model training.

The second layer in *word encoder* is a convolutional neural network (CNN) [15]. A search query is usually a combination of several keywords, such as "peppa pig website", rather than a complete sentence [22]. Thus, it may be not suitable to regard a query as a word sequence and use sequence modeling methods such as LSTM to learn the representation of search query. In addition, the local contexts of words are very important for learning query representations. For example, combined with "peppa" we know "pig" is a part of a cartoon name which is informative for age prediction. However, in other contexts such as "domestic pig" this word is not so informative for age prediction. Thus, we propose to learn query representations by using CNN to capture the local contextual information of words. Denote $\mathbf{f}_w \in \mathcal{R}^{K_w D}$ as a filter in the CNN network for word encoder with window size K_w, then the contextual representation of the i-th word in the query q learned by this filter is formulated as follows:

$$c_i = g(\mathbf{f}_w{}^T \times \mathbf{w}_{\lfloor i-\frac{K_w-1}{2}\rfloor:\lfloor i+\frac{K_w-1}{2}\rfloor} + b_w), \quad (1)$$

where $\mathbf{w}_{\lfloor i-\frac{K_w-1}{2}\rfloor:\lfloor i+\frac{K_w-1}{2}\rfloor}$ is the combination of the embedding vectors of words from position $\lfloor i - \frac{K_w-1}{2}\rfloor$ to position $\lfloor i + \frac{K_w-1}{2}\rfloor$, and g is the activation function which is ReLU [15] in our approach. We use multiple filters in the CNN layer and the final contextual representation of the i-th word is the concatenation of the outputs of these filters at position i, which is denoted as $\mathbf{c}_i \in \mathcal{R}^{F_w}$, where F_w is the number of filters.

The third layer in *word encoder* is an attention network [17]. Different words in the same search query may have different importance for demographic prediction. For example, in the query "christmas gift for girlfriend", the word "girlfriend" may be more informative than "gift" in inferring user's gender. In addition, the same word may have different informativeness in different search queries. For instance, the word "girlfriend" in the query "film the new girlfriend" is less informative than in aforementioned query. It is important to select useful words in different contexts to learn more informative query representations. Thus, we use a word-level attention network to help our model attend differently to different words in different contexts according to their importance to demographic prediction. Following many previous works on attention mechanism [39], the attention weight of word w_i in query q is

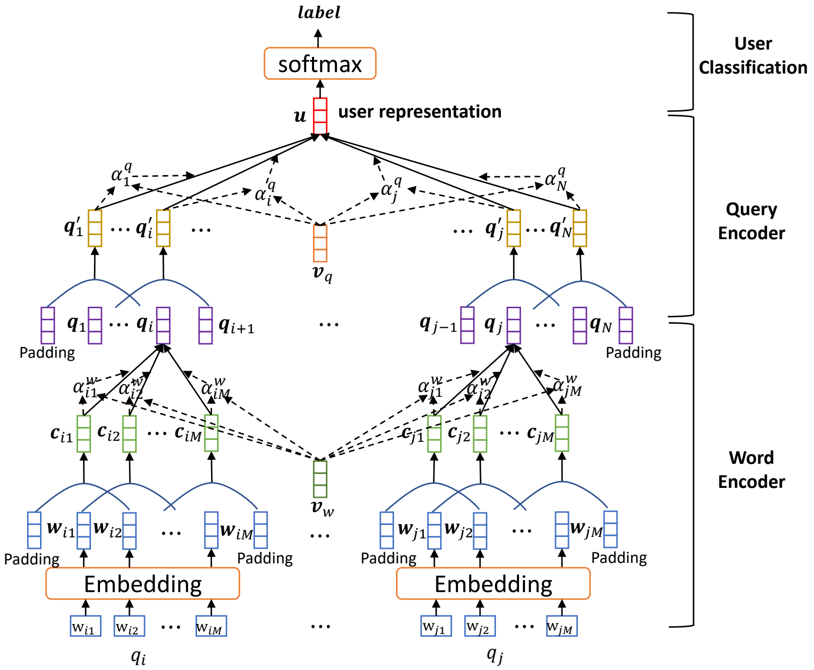

label
softmax
u
user representation
User Classification
α_1^q
α_i^q
α_j^q
α_N^q
q'_1 ··· q'_i ···
··· q'_j ··· q'_N
v_q
Query Encoder
Padding
q_1 ··· q_i q_{i+1}
···
q_{j-1} q_j ··· q_N
Padding
α_{i1}^w α_{i2}^w α_{iM}^w
α_{j1}^w α_{j2}^w α_{jM}^w
c_{i1} c_{i2} ··· c_{iM}
c_{j1} c_{j2} ··· c_{jM}
v_w
Word Encoder
Padding w_{i1} w_{i2} ··· w_{iM} Padding
···
Padding w_{j1} w_{j2} ··· w_{jM} Padding
Embedding
Embedding
w_{i1} w_{i2} ··· w_{iM}
···
w_{j1} w_{j2} ··· w_{jM}
q_i
q_j

Figure 3: The overall framework of our HURA approach.

computed as follows:

$$\mathbf{v}_i^w = \tanh(\mathbf{V}_w{}^T \times \mathbf{c}_i + \mathbf{b}_w), \tag{2}$$

$$\alpha_i^w = \frac{\exp(\mathbf{v}_w{}^T \times \mathbf{v}_i^w)}{\sum_{j=1}^{M} \exp(\mathbf{v}_w{}^T \times \mathbf{v}_j^w)}, \tag{3}$$

where $\mathbf{V}_w \in \mathcal{R}^{F_w \times T_w}$, $\mathbf{v}_w \in \mathcal{R}^{T_w}$ and $\mathbf{b}_w \in \mathcal{R}^{T_w}$ are the parameters of the word-level attention network, and α_i^w is the contextual attention weight of word w_i in query q. $\mathbf{v}_w$ is usually called "context vector" [39], and can encode the information of "what words are important for demographic prediction in a context".

The final representation of search query q is the summation of the contextual representations of its words weighted by their attention weights, which is formulated as follows:

$$\mathbf{q} = \sum_{i=1}^{M} \alpha_i^w \times \mathbf{c}_i. \tag{4}$$

3.2 Query Encoder

The *query encoder* module in our HURA approach is used to learn the representation of a user u from his/her queries $[q_1, q_2, ..., q_N]$. As illustrated in Fig. 3, this module contains two layers.

The first layer is a query-level CNN network to learn the contextual representations of search queries by capturing their local contexts. The search queries posted by the same user in neighboring time may have relatedness with each other [5, 22, 34]. For example, they may be the different formulations of the same search intent or on different aspects of the same search target. The relatedness between neighbouring queries is useful for learning more accurate and robust representations of search queries, since the context information in a singe query is usually limited and the related neighbouring queries can provide complementary information. However, the relatedness between queries with long time span is usually very weak. Thus, we use CNN network to capture the local contexts of search queries to enhance the learning of query representations. The input of this CNN network is the representation vectors of user u's search queries $[\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_N]$ which are learned by the *word encoder* module. Denote $\mathbf{f}_q \in \mathcal{R}^{K_q F_w}$ as a filter in the query-level CNN network with window size K_q, then the contextual representation of query q_i learned by this filter is computed as follows:

$$q'_i = g(\mathbf{f}_q{}^T \times \mathbf{q}_{\lfloor i-\frac{K_q-1}{2} \rfloor : \lfloor i+\frac{K_q-1}{2} \rfloor} + b_q), \tag{5}$$

where $\mathbf{q}_{\lfloor i-\frac{K_q-1}{2} \rfloor : \lfloor i+\frac{K_q-1}{2} \rfloor}$ is the combination of representation vectors of search queries from $\lfloor i - \frac{K_q-1}{2} \rfloor$ to $\lfloor i + \frac{K_q-1}{2} \rfloor$, and g is the ReLU activation function. The final contextual representation of search query q_i is a concatenation of the outputs of multiple filters, denoted as $\mathbf{q}'_i \in \mathcal{R}^{F_q}$, where F_q is the number of filters in the query-level CNN network.

The second layer is a query-level attention network. Different search queries contribute differently to the demographic prediction, and many search queries are irrelevant and even noisy for this task. For example, search queries like "birthday gift for grandson" and "philips shaver" may be more informative than queries like "google maps" and "amazon.com" for age and gender prediction. Thus, we use a query-level attention network to help our model attend differently to different search queries to learn more informative user representations from search queries. The attention weight of search query q_i is computed as follows:

$$\mathbf{v}_i^q = \tanh(\mathbf{V}_q{}^T \times \mathbf{q}'_i + \mathbf{b}_q), \tag{6}$$

$$\alpha_i^q = \frac{\exp(\mathbf{v}_q{}^T \times \mathbf{v}_i^q)}{\sum_{j=1}^{N} \exp(\mathbf{v}_q{}^T \times \mathbf{v}_j^q)}, \tag{7}$$

where $\mathbf{V}_q \in \mathcal{R}^{F_q \times T_q}$, $\mathbf{v}_q \in \mathcal{R}^{T_q}$ and $\mathbf{b}_q \in \mathcal{R}^{T_q}$ are the parameters of the query-level attention network, and α_i^q is the attention weight of query q_i. The context vector $\mathbf{v}_q$ is used to encode the information of what search queries are important for a specific user demographic prediction task, e.g., age prediction.

The final hidden representation of user u is the summation of the contextual representations of his/her search queries weighted by attention weights of these queries, which is formulated as follows:

$$\mathbf{u} = \sum_{i=1}^{N} \alpha_i^q \times \mathbf{q}_i'. \tag{8}$$

3.3 User Classification

The *user classification* module is used to classify a user into one of the predefined demographic categories (e.g., an age group in age prediction and a gender category in gender prediction) according to the hidden representation of this user learned from his/her search queries. In this module, we use a softmax layer to compute the probabilities of user u in different demographic categories, which is formulated as follows:

$$\mathbf{p}_u = \text{softmax}(\mathbf{W}^T \times \mathbf{u} + \mathbf{b}_u), \tag{9}$$

where $\mathbf{W} \in \mathcal{R}^{F_w \times C}$ and $\mathbf{b}_u \in \mathcal{R}^C$ are the parameters of the *user classification* layer, and C is the number of categories.

In the model training stage, we use crossentropy as the loss function, and the overall objective function is formulated as follows:

$$\mathcal{L} = -\sum_{u=1}^{U} \sum_{c=1}^{C} y_{u,c} \log p_{u,c}, \tag{10}$$

where $y_{u,c}$ is the gold label of user u in category c, which is 1 if c is the true and 0 otherwise. U is the number of users for training.

In the prediction stage, the label with the largest score in $\mathbf{p}_u$ is selected as the predicted demographic category for user u.

4 EXPERIMENTS

4.1 Datasets and Experimental Settings

Since there is no off-the-shelf datasets for search query based demographic prediction, we built two datasets by ourselves. The first dataset is for age prediction (denoted as *Age*). We randomly sampled 25,000 users from a commercial search engine and the age categories of these users are included in their profiles. We also downloaded the search queries of these users in 6 months, ranging from October 1, 2017 to March 31, 2018. The average number of search queries per user is 171.6, and the average query length is 3.42 words. Each user has an age category, and the mapping between age category and age range is summarized in Table 1. The distribution of user numbers in different age categories is shown in Fig. 4. The second dataset is for gender prediction (denoted as *Gender*). We also randomly sampled 25,000 users from a commercial search engine whose gender information is available and downloaded their search queries in the same way as the *Age* dataset. There are 13,349 male users and 11,651 female users in this dataset. In both datasets we randomly sampled 20,000 users as training set, and the remaining users as test set. Among the training users we randomly sampled 10% of them for validation.

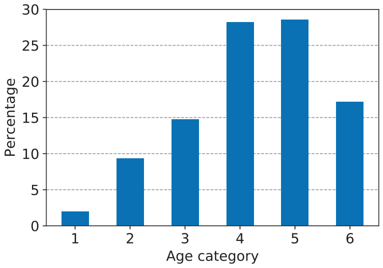


Figure 4: The distribution of users in different age groups.

Table 1: The mapping between age category and age range.

Age category	1	2	3	4	5	6
Age range	< 18	[18, 24]	[25, 34]	[35, 49]	[50, 64]	> 64

In our experiments, the word embeddings were pretrained on all the search queries in the training data using the word2vec[1] tool. The embedding dimension is 200 for both datasets. The window sizes of word-level CNN and query-level CNN are both 3, and the number of filters in these CNN networks is 300. RMSProp [4] was used as the optimizer for model training. The batch size was set to 100. T_w in the word-level attention network and T_q in the query-level attention network are both set to 300. Dropout technique [35] was used in our approach to mitigate overfitting and was applied to both word embedding layer and CNN layers. The dropout rate was set to 0.2. Early stopping strategy [2] was also used. If the loss on validation data does not decline after 3 epochs, then the training will be terminated. All the hyperparameters were selected according to the validation data. Each experiment was repeated for 10 times independently and average results were reported.

4.2 Performance Evaluation

In this section we evaluate the performance of our HURA approach by comparing it with several baseline methods for user demographic prediction. The methods to be compared include:

- *SVM*: support vector machine, widely used in demographic prediction [9].
- *LR*: logistic regression, also very popular for demographic prediction [25, 32].
- *LinReg*: linear regression with Lasso regularization [26].
- *FastText*: a popular method for text classification [13].
- *CNN*: demographic prediction based on convolutional neural network [14].
- *LSTM*: demographic prediction based on Long Short-Term Memory network [11].
- *HAN*: a hierarchical attention network for document classification [39]. We regard each user a document and each search query as a sentence.
- *HURA*: our proposed approach for search query based user demographic prediction.

[1] https://code.google.com/archive/p/word2vec/

Table 2: Experimental results on the *Age* dataset. *Fscore* is macro-averaged Fscore.

	10%		50%		100%	
	Accuracy	Fscore	Accuracy	Fscore	Accuracy	Fscore
SVM	32.92±0.26	31.88±0.24	36.33±0.32	35.74±0.32	39.59±0.40	39.10±0.39
LR	34.25±0.27	33.55±0.25	37.70±0.33	37.18±0.32	40.68±0.44	39.92±0.43
LinReg	31.45±0.29	30.66±0.27	34.10±0.36	33.29±0.35	37.34±0.54	36.15±0.53
FastText	32.56±0.29	32.04±0.28	35.94±0.35	35.33±0.34	39.32±0.37	38.86±0.36
CNN	35.68±0.72	34.89±0.69	38.72±0.63	38.25±0.58	41.34±0.59	40.64±0.54
LSTM	34.97±0.68	34.12±0.64	38.29±0.61	37.63±0.59	40.73±0.58	40.08±0.54
HAN	36.89±0.66	36.14±0.62	40.04±0.56	39.25±0.53	42.17±0.57	41.64±0.52
HURA	38.57±0.62	37.94±0.58	42.58±0.55	41.75±0.50	44.42±0.54	42.58±0.48

Table 3: Experimental results on the *Gender* dataset. *Fscore* is macro-averaged Fscore.

	10%		50%		100%	
	Accuracy	Fscore	Accuracy	Fscore	Accuracy	Fscore
SVM	60.76±0.14	60.43±0.14	62.40±0.16	62.32±0.16	63.53±0.18	63.51±0.19
LR	61.43±0.16	61.40±0.16	62.93±0.17	62.91±0.18	63.67±0.21	63.65±0.22
LinReg	57.49±0.20	56.86±0.20	60.05±0.23	59.92±0.23	61.47±0.24	61.42±0.24
FastText	60.79±0.18	60.66±0.18	61.94±0.17	61.90±0.18	63.39±0.16	63.34±0.16
CNN	64.45±0.44	64.39±0.46	67.21±0.31	67.15±0.32	68.66±0.28	68.63±0.28
LSTM	64.02±0.39	63.96±0.40	66.74±0.30	66.69±0.30	68.43±0.27	68.38±0.27
HAN	64.94±0.37	64.88±0.38	68.37±0.29	68.31±0.30	69.85±0.26	69.80±0.27
HURA	66.70±0.34	66.68±0.35	69.94±0.27	69.91±0.27	71.14±0.25	71.12±0.26

Following [25], the features used in SVM, LR and LinReg are word unigrams. The hyperparameters of these baseline methods were tuned on validation data. We conducted experiments on different ratios of training data (i.e., 10%, 50% and 100%) to explore the performance different methods with different amounts of labeled data. Classification accuracy and macro-averaged Fscore are used as evaluation metrics. The results are summarized in Tables 2 and 3 .

According to Tables 2 and 3, our HURA approach achieves the best performance among all the methods compared here on both age prediction and gender prediction tasks. The hypothesis testing results show that HURA can significantly outperform baseline methods at the significance level of 0.01 evaluated by t-test. Our HURA approach can perform better than many popular user demographic prediction methods that are based on traditional machine learning methods, such as SVM [9], LR [32], and LinReg [26]. This is because the features for representing users in these methods are handcrafted, and cannot capture the complex semantics and contexts of search queries. In our approach, the feature vector for user representation is learned from data using a neural network based model, which is more suitable for search query based demographic prediction. Our approach can also outperform many existing demographic prediction methods that are based on neural networks, such as FastText [13], CNN [14] and LSTM [11]. In these methods, all the search queries of the same user are concatenated together as a long text for building user representation. Thus, the semantic information of each individual search query cannot be effectively captured, and these methods cannot distinguish useful search queries from less useful queries. Since many search queries are uninformative and even noisy for demographic prediction, these methods may be suboptimal for search query based age and gender prediction. Different from these deep learning based demographic prediction methods, in our HURA approach we use a hierarchical neural model for user representation which first learns a representation vector for each query from their words using a word encoder, and then learns a representation vector for each user based on their search queries using a query encoder. In addition, we incorporate both word-level and query-level attention networks into our HURA approach to attend differently to different words and queries based on their contributions to demographic prediction task. Since different words and different search queries have different informativeness for demographic prediction, our approach is more appropriate for predicting users' ages and genders from their search queries than these existing deep learning based demographic prediction methods, which is validated by the experimental results. Although HAN [39] can exploit the hierarchical structure of document (i.e., words form sentences and sentences form documents), our HURA approach performs consistently better than it. This is because in the HAN method LSTM network is used to learn sentence representation from words and learn document representation from sentences. This method is appropriate for document classification and the sequential information of words and sentences can be captured. However, search queries are quite different from sentences, since a

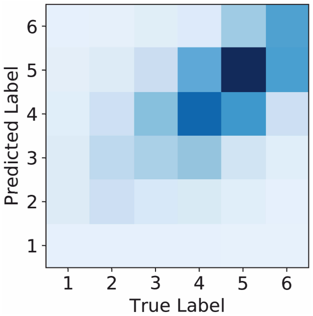
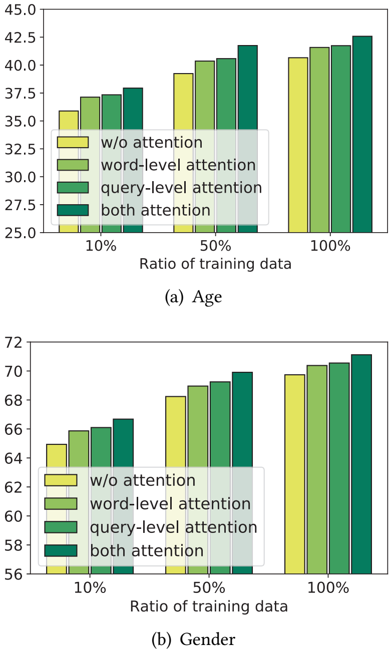


Figure 5: The confusion matrix of our approach on the *Age* dataset. Deeper color represents larger values.



(a) Age

(b) Gender

Figure 6: The effectiveness of word-level and query-level attention networks for our HURA approach.

search query is usually a combination of several keywords, rather than a complete sentence. In addition, different from documents where sentences in the same document usually have strong relatedness with each other (e.g., following the same topic), there is more diversity among the search queries from the same user. For example, a user may search for various kinds of information in different search sessions. Although search queries in neighboring time may have relatedness with each other, the relatedness between search queries with a long time span is usually very weak. Thus, capturing the local contexts of words and search queries using CNN may be more appropriate for query representation and user representation. Thus, our HURA approach can consistently outperform HAN in the search query based demographic prediction task.

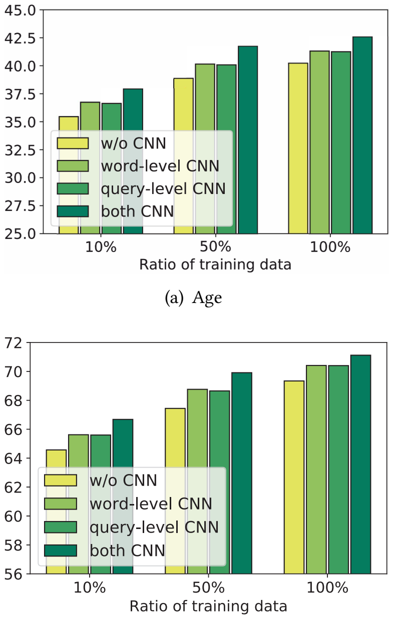
In order to further evaluate the performance of our approach, we calculated the confusion matrix of our approach on the test data. The results on the *Age* dataset are shown in Fig. 5. We can see that most of the misclassifications are caused by classifying users into neighboring age categories. For example, many users with age category 4 (ages in [35, 49]) are misclassified into age category 5 (ages in [50, 64]). This result is consistent with our intuition, since there are many overlaps in searching contents and writing styles between users in neighboring age categories. Thus, the experimental results show that the prediction results of our approach are reasonable.

4.3 Model Effectiveness

In this section we conducted several experiments to explore the effectiveness of the model of our HURA approach. First, we want to verify whether the word-level and query-level attention networks are useful for improving the performance of our approach. The experimental results on both datasets are summarized in Fig. 6.

According to Fig. 6, incorporating the word-level attention network can effectively improve the performance of our HURA approach. This is because different words usually have different importance in building informative query representation for demographic prediction. For example, in the query "birthday gift for wife" the word "wife" is more informative than "gift" in inferring the gender of this user. Thus, the word-level attention network can help our model attend differently to different words according to their contributions to demographic prediction. In addition, incorporating the query-level attention network can also improve the performance of HURA. This is due to that different search queries have different informativeness for predicting demographics of users. For example, the query "birthday gift for wife" is more informative than "gmail login" for gender prediction, and the query "gift for classmates" is more informative than "amazon.com" for age prediction. Thus, the query-level attention network is very useful for our HURA approach to recognize and highlight informative search queries to learn more informative user representations for demographic prediction. Moreover, incorporating both word-level and query-level attention networks can further improve the performance of our HURA approach, which indicates that they are complementary with each other.

We also conducted experiments to explore the effectiveness of word-level and query-level CNN networks in learning contextual representations of words and search queries for demographic prediction. The experimental results are summarized in Fig. 7. We can see that incorporating the word-level CNN network can effectively improve the performance of our approach. This is because neighboring words in search queries usually have dependencies with each other, e.g., belonging to the same phrase or entity name, and capturing the local contexts of words is beneficial for learning high-quality query representations. In addition, incorporating the query-level CNN network is also helpful. This is because neighboring queries from the same user may have relatedness with each other. For example, they may be the different formulations of the same search intent. Thus, capturing the local context information of search queries can help learn more accurate and robust query representations, since search queries are usually very short and the context information in individual queries is limited. Moreover, by incorporating both word-level and query-level CNN networks our approach can achieve better performance.



(a) Age

(b) Gender

Figure 7: The effectiveness of word-level and query-level CNN networks for our HURA approach.

4.4 Case Study

In this section, we conducted several case studies to further explore how our approach works. More specifically, we want to verify whether the word-level and query-level attention networks can select important words and search queries for demographic prediction. In Figs. 8 and 9 we show several example search queries from randomly selected users in the test data of the *Age* and *Gender* datasets.

From Figs. 8 and 9, we have several observations. First, the search queries are very short and usually contain only a few words. Thus, the context information is limited in each single search query. Second, search queries from the same user can be very diverse in content, and many of them are not useful for inferring demographics, such as "signin", "amazon' and "mail". Thus, predicting users' demographics based on their search queries is very challenging. Third, neighboring words in the same query usually have dependencies with each other, e.g., the words "pioneer" and "woman" in the query "pioneer woman magazine". In addition, the search queries posted by the same user in neighboring time may also have relatedness. For example, "elderly tax credit form" and "county elderly tax credit form" are two neighboring queries posted by the same user. These two search queries are highly related to each other and they share the same search intent. Thus, capturing the local contexts of words and search queries is beneficial for learning more informative user representations to predict user demographics.

In addition, according to Figs. 8 and 9, our HURA approach can distinguish useful words from less useful words using the word-level attention network. For example, in Fig. 8(a) the words "cool", "games" and "quiz" are assigned high attention weights. From these words we can infer that this user is probably a teenager. However, words like "mail" and "login" are assigned low attention weights, since they are widely used by all users and are not informative enough. In Fig. 9, the words "lipstick" and "skirt" are assigned higher attention weights than "gift" and "element", since "lipstick" and "skirt" are more informative for gender prediction. In addition, our HURA model can distinguish informative search queries from less informative queries using the query-level attention network. For example, in Fig. 9 search queries likes "give my wife a gift" and "best electric shaver" are assigned high attention weights since they are informative for gender prediction, while queries like "youtube.com" and "amazon" are assigned low attention weights since they are frequently used by both male and female users. Thus, these case studies validate that the motivations of our approach are reasonable and our HURA approach is effective in selecting important information for predicting demographics based on search queries.

5 CONCLUSION

In this paper we study an interesting and challenging problem, i.e., predicting the demographics of online users based on their search queries. We propose a neural approach named HURA for this task. The core of our approach is a hierarchical neural model to learn user representations from search queries for demographic prediction. Our model first uses a word encoder to learn representations of queries from words, and then uses a query encoder to learn user representations from queries. In addition, we incorporate both word-level and query-level attention networks into our approach to select and highlight important words and search queries to learn more informative user representations for demographic prediction. Experiments on two real-world datasets show that our approach can effectively improve the performance of search query based age and gender prediction, and consistently outperform many baseline methods.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions that helped improve the quality of this paper. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0800402 and in part by the National Natural Science Foundation of China under Grant U1705261, Grant U1536207, Grant U1536201 and U1636113.

REFERENCES

[1] Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. 2013. Inferring the demographics of search users: Social data meets search queries. In *WWW*. 131–140.

[2] Rich Caruana, Steve Lawrence, and C Lee Giles. 2001. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *NIPS*. 402–408.

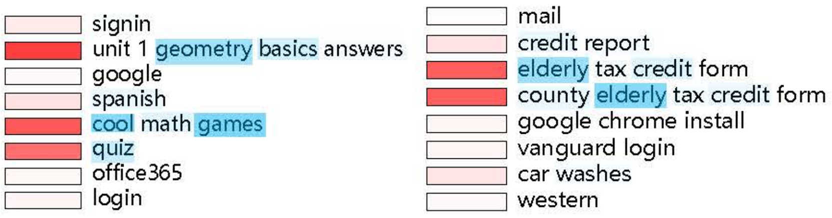
[3] Aron Culotta, Nirmal Kumar Ravi, and Jennifer Cutler. 2016. Predicting Twitter user demographics using distant supervision from website traffic data. *Journal of Artificial Intelligence Research* 55 (2016), 389–408.

[4] Yann Dauphin, Harm de Vries, and Yoshua Bengio. 2015. Equilibrated adaptive learning rates for non-convex optimization. In *NIPS*. 1504–1512.

[5] Mostafa Dehghani, Sascha Rothe, Enrique Alfonseca, and Pascal Fleury. 2017. Learning to attend, copy, and generate for session-based query suggestion. In *CIKM*. ACM, 1747–1756.

[6] Golnoosh Farnadi, Jie Tang, Martine De Cock, and Marie-Francine Moens. 2018. User Profiling through Deep Multimodal Fusion. In *WSDM*. 171–179.

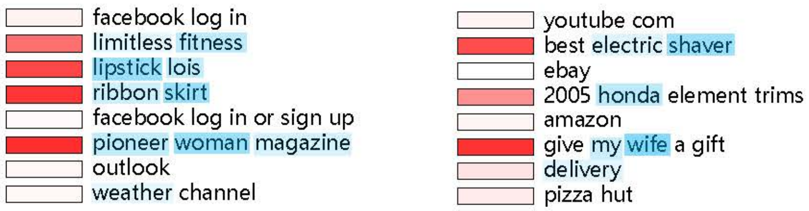
[7] Katja Filippova. 2012. User demographics and language in an implicit social network. In *EMNLP*. 1478–1488.



(a) Queries from a teenager user.

(b) Queries from an elder user.

Figure 8: Example search queries from two randomly selected users in the *Age* dataset. Red bars and blue bars represent query-level and word-level attention weights respectively. Deeper color represents higher attention weights.



(a) Queries from a female user.

(b) Queries from a male user.

Figure 9: Example search queries from two randomly selected users in the *Gender* dataset.

[8] Clayton Fink, Jonathon Kopecky, and Maksym Morawski. 2012. Inferring Gender from the Content of Tweets: A Region Specific Example.. In *ICWSM*. 459–462.

[9] Sharad Goel, Jake M Hofman, and M Irmak Sirer. 2012. Who Does What on the Web: A Large-Scale Study of Browsing Behavior.. In *ICWSM*. 120–137.

[10] Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric Analysis of Bloggers' Age and Gender. In *ICWSM*. 214–217.

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[12] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. 2007. Demographic prediction based on user's browsing behavior. In *WWW*. 151–160.

[13] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *EACL*, Vol. 2. 427–431.

[14] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP*. 1746–1751.

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436.

[16] Wen Li and Markus Dickinson. 2017. Gender Prediction for Chinese Social Media Data. In *Proc. of Recent Advances in Natural Language Processing*. 438–445.

[17] Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *EMNLP*. 1412–1421.

[18] Sunghwan Mac Kim, Qiongkai Xu, Lizhen Qu, Stephen Wan, and Cécile Paris. 2017. Demographic Inference on Twitter using Recursive Neural Networks. In *ACL*, Vol. 2. 471–477.

[19] Ian MacKinnon and Robert H Warren. 2007. Age and geographic inferences of the LiveJournal social network. In *Statistical Network Analysis: Models, Issues, and New Directions*. Springer, 176–178.

[20] Eric Malmi and Ingmar Weber. 2016. You Are What Apps You Use: Demographic Prediction Based on User's Apps.. In *ICWSM*. 635–638.

[21] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the Demographics of Twitter Users. *ICWSM* 11, 5th (2011), 25.

[22] Bhaskar Mitra. 2015. Exploring session context using distributed representations of queries and reformulations. In *SIGIR*. 3–12.

[23] Saif M Mohammad and Tony Wenda Yang. 2011. Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*. 70–79.

[24] Antonio A Morgan-Lopez, Annice E Kim, Robert F Chew, and Paul Ruddle. 2017. Predicting age groups of Twitter users based on language and metadata features. *PloS one* 12, 8 (2017), e0183537.

[25] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. " How Old Do You Think I Am?" A Study of Language and Age in Twitter.. In *ICWSM*. 439–448.

[26] Dong Nguyen, Noah A Smith, and Carolyn P Rosé. 2011. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. 115–123.

[27] Dong Nguyen, Dolf Trieschnigg, A Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska De Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *COLING*. 1950–1961.

[28] Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents*. 37–44.

[29] Bryan Perozzi and Steven Skiena. 2015. Exact age prediction in social networks. In *WWW*. 91–92.

[30] Zhen Qin, Yilei Wang, Yong Xia, Hongrong Cheng, Yingjie Zhou, Zhengguo Sheng, and Victor CM Leung. 2014. Demographic information prediction based on smartphone application usage. In *2014 International Conference on Smart Computing*. IEEE, 183–190.

[31] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*. 37–44.

[32] Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *ACL*. 763–772.

[33] K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma. 2013. Author profiling: Predicting age and gender from blogs. *Notebook for PAN at CLEF* (2013), 119–124.

[34] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *CIKM*. ACM, 553–562.

[35] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[36] Jingjing Wang, Shoushan Li, and Guodong Zhou. 2017. Joint learning on relevant user attributes in micro-blog. In *IJCAI*. 4130–4136.

[37] Liang Wang, Qi Li, Xuan Chen, and Sujian Li. 2016. Multi-task Learning for Gender and Age Prediction on Chinese Microblog. (2016), 189–200.

[38] Ingmar Weber and Carlos Castillo. 2010. The demographics of web search. In *SIGIR*. 523–530.

[39] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL*. 1480–1489.

[40] Josh Jia-Ching Ying, Yao-Jen Chang, Chi-Min Huang, and Vincent S Tseng. 2012. Demographic prediction based on users mobile behaviors. *Mobile Data Challenge* (2012), 1–6.

[41] Dong Zhang, Shoushan Li, Hongling Wang, and Guodong Zhou. 2016. User classification with multiple textual perspectives. In *COLING*. 2112–2121.