You Are Where You Go: Inferring Demographic Attributes from Location Check-ins

Yuan Zhong[†‡], Nicholas Jing Yuan[†], Wen Zhong[†⋆], Fuzheng Zhang[§†], Xing Xie[†]
[†]Microsoft Research
[‡]Northeastern University, Boston, MA, USA
[⋆]Stony Brook University, Stony Brook, NY, USA
[§] University of Science and Technology of China, Hefei, China
{nicholas.yuan, xingx} at microsoft.com, zhong.yu at husky.neu.edu, wen.zhong at stonybrook.edu, zhfzh at mail.ustc.edu.cn

ABSTRACT

User profiling is crucial to many online services. Several recent studies suggest that demographic attributes are predictable from different online behavioral data, such as users' "Likes" on Facebook, friendship relations, and the linguistic characteristics of tweets. But location check-ins, as a bridge of users' offline and online lives, have by and large been overlooked in inferring user profiles.

In this paper, we investigate the predictive power of location check-ins for inferring users' demographics and propose a simple yet general *location to profile (L2P)* framework. More specifically, we extract rich semantics of users' check-ins in terms of *spatiality*, *temporality*, and *location knowledge*, where the location knowledge is enriched with semantics mined from heterogeneous domains including both online customer review sites and social networks. Additionally, tensor factorization is employed to draw out low dimensional representations of users' intrinsic check-in preferences considering the above factors. Meanwhile, the extracted features are used to train predictive models for inferring various demographic attributes.

We collect a large dataset consisting of profiles of $159,530$ verified users from an online social network. Extensive experimental results based upon this dataset validate that: 1) Location check-ins are diagnostic representations of a variety of demographic attributes, such as gender, age, education background, and marital status; 2) The proposed framework substantially outperforms compared models for profile inference in terms of various evaluation metrics, such as precision, recall, F-measure, and AUC.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; D.4.8 [**Performance**]: Modeling and prediction; J.4 [**Social and Behavioral Science**]: Sociology

General Terms

Algorithms, Performance, Measurement

WSDM'15, February 2–6, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2684822.2685287.

Keywords

Demographics; spatiality; temporality; location knowledge; tensor facotorization; prediction

1. INTRODUCTION

Long conceived as a significant research area, the inference of demographic attributes has been extensively studied in linguistics [14, 26], psychology [8], and sociology [17]. Before the rise of social networks, a host of different works have looked into data sources from a myriad of domains, such as Internet browsing behaviors [18], written texts [11], telephone conversations [5], real-world mobile network communication records (CALL and SMS) [10] and answers to specific psychological tests [8] to predict various profile attributes like gender, age, income, and personality.

The recent explosive growth of online social media and social networking websites has allowed billions of users around the world to create and view friendship connections, self-expressions, newsfeeds, etc., and lifted the role of profile attributes to an even higher position. Social networks like Facebook[1], Sina Weibo[2], and Renren[3] treat certain profiles (gender, birthday, residence, etc.) as mandatory information for registration. With these attributes, users are served with a personalized experience. Existing works have examined diverse data sources on social networks, validating that communication behaviors (replying/retweeting [23]/"Likes" [13]), linguistic characteristics of tweets [20], group memberships [16] and friendships [33] are useful for profile inference.

However, *human mobility*, as a very informative and fundamental user behavior, has been overlooked by most previous works for profile inference. In fact, users' demographic attributes are of great importance in terms of at least two conspicuous aspects: 1) commerce: the profile contributes significantly to link prediction, item recommendation, and targeted advertising, which are crucial for most companies; 2) uses: the profile is directive for content sharing, membership attachment, and trust establishment, which are pervasive for various personalized services.

In this paper, we look into demographics inference through one of the most popular human mobility data, location check-ins from online social networks. In particular, we uncover three distinct aspects embodied in the check-in data that are potentially correlated with users' profiles, namely *temporality*, *spatiality* and *location knowledge*.

[1] http://www.facebook.com
[2] http://www.weibo.com/signup/signup.php
[3] http://www.renren.com

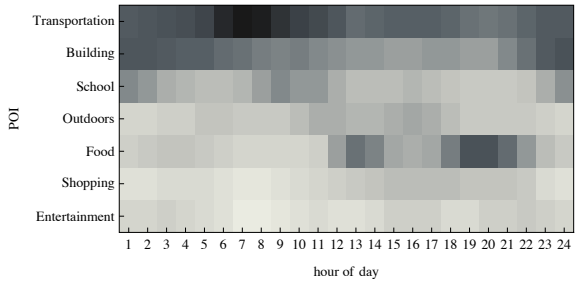


Figure 1: Hourly density distribution of check-ins w.r.t different categories of POIs

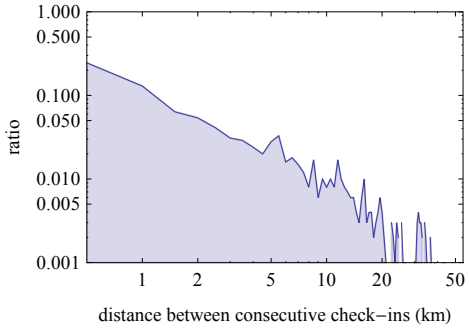
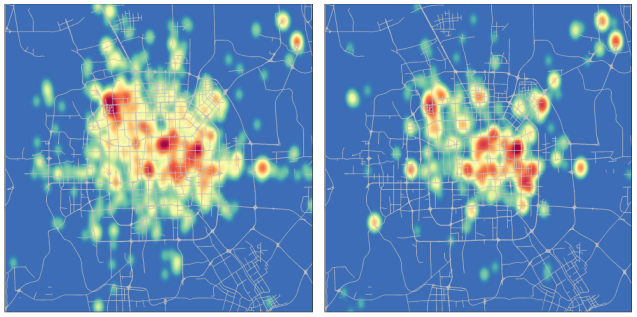


Figure 2: Distance between consecutive check-ins



(a) native check-ins in BJ (b) non-native check-ins in BJ

(c) native check-ins in SH (d) non-native check-ins in SH

Figure 3: Density distribution of native and nonnative check-ins in Beijing (BJ) and Shanghai (SH)

• **Temporality**: Human mobility is imbued with ample temporal patterns at different granularities, e.g., day of the week and time of day. For example, office staff commute from home to their company every weekday morning. It's common to see a retired person shopping in the supermarket on a weekday afternoon and a taxi driver working at midnight during holidays. Recent studies have also found that human mobility follows a high degree of regularity [12]. As an example, Figure 1 shows the hourly density distribution of different types of check-ins from Sina Weibo (China's Twitter) based on 10,000 randomly sampled users' check-in histories, which clearly indicates different patterns for various categories of POIs, e.g., the concentration of "food" related check-ins mostly occur at noon and dinner time, while most transportation related check-ins occur in the early morning.

• **Spatiality**: Human mobility is the manifestation of spatial variation. More specifically, the spacial scale is restricted by transportation conditions. For instance, one cannot check in at Stanford University if he has just checked in at Northeastern University only half an hour before. This makes mobility distinctive from unrestrained behaviors such as "Liking", following, replying and retweeting on social networks. For example, Figure 2 presents the distance between consecutive check-ins of 100,000 randomly sampled users, in which we find that a large proportion of the distance is less than 20 kilometers. As another example, we plot the density distributions of check-ins by natives and non-natives in Beijing and Shanghai. As is shown in Figure 3, natives' check-ins (Figure 3(a) and Figure 3(c)) are more diverse in terms of geo-spatial distribution while the non-natives check-ins (Figure 3(b) and Figure 3(d)) are skewed to certain scenic areas and hot spots.

• **Location knowledge**: Human mobility strongly correlates to the functionality of locations which motivates people to travel between different places. Apt instances are: students go to school because they acquire knowledge there; businessmen go to a city's central business district because they conduct commercial affairs there; people go to restaurant districts because they have lunch or dinner there. A check-in is typically associated with a Point of Interest (POI), which belongs to a certain category, e.g., teaching building or shopping mall. Furthermore, the semantics of a location sometimes contain far more information than just the category, e.g., the atmosphere and price range of a restaurant, or the quality of a college.TThese semantics can be enriched with "human knowledge" which are revealed from customer review sites (such as Yelp[4] and Dianping[5]) and online social networks (when users mention these locations).

Based on these factors, we propose a location to profile *(L2P)* framework for inferring demographic attributes of online users including gender, age, education background, sexual orientation, marital status, blood type and zodiac sign. To the best of our knowledge, this is the first work focusing on inference of demographic attributes through location check-ins. On the whole, this paper offers the followings contributions:

• We exploit the semantics of user location check-ins from three points of views, i.e. *spatiality*, *temporality* and *location knowledge*. In particular, we conflate Points of Interest (POIs) of Sina Weibo with Dianping (a review site known as China's Yelp) and explicitly learn the location knowledge of a POI from various aspects such as *categories*, *reviews* and *keywords*.

• We propose an *L2P* framework, incorporating *spatiality*, *temporality*, *location knowledge* features with a tensor model so as to in-

[4] http://www.yelp.com
[5] http://www.dianping.com

Table 1: Demographic Attributes and Corresponding Categories

Attribute	Completion Rate	Categories
gender	94.0193%	male, female
age	33.1588%	the specific age number
education background	36.7228%	university, non-university
sexual orientation	2.5549%	heterosexuality, bisexuality, male homosexuality, female homosexuality
marital status	2.6396%	single, courtship (seeking a relationship), in love, married
blood type	1.6376%	O, A, B, AB
zodiac sign	58.1649%	twelve zodiac signs

fer various demographics (e.g. sexual orientation, education background and marital status) for a particular user based on check-in records.

• Based on a large dataset collected from Sina Weibo, we have conducted extensive experiments to validate the effectiveness of the proposed feature and framework. The results show that our approach significantly outperforms baseline models in terms of multiple metrics such as precision, recall, F-measure and area under curve (AUC).

The rest of this paper is structured as follows: Section 2 introduces the experimental dataset and demographic attributes. The *L2P* framework is proposed in Section 3. Experimental results based on a large-scale dataset are presented in Section 4. Related work is reviewed in Section 5 and the paper is concluded in Section 6 with a brief discussion of limitations and directions of future research.

Table 2: Demographics Distribution

(a) Marital Status

Status Type	Fraction
single	59.31%
courtship	17.07%
in love	12.74%
married	10.87%

(b) Blood Type

Blood Type	Fraction
O	34.97%
A	25.36%
B	25.66%
AB	14.00%

(c) Sexual Orientation

Orientation Type	Fraction
heterosexuality	86.26%
male homosexuality	2.66%
female homosexuality	1.68%
bisexuality	9.40%

2. DATASET AND DEMOGRAPHICS

In this section, we introduce the collected dataset of user profiles and location check-ins, and devise a method for generating ground truth data that is used to train and validate the performance of our inference model.

2.1 Dataset

In order to investigate and evaluate the performance of our inference framework, we crawled 3,354,918 users' profiles on Sina Weibo whose residential cities are provided as "Beijing" or "Shanghai" (two metropolitan cities in China) from Sina Weibo using the LifeSpec data platform [32]. The profile information also contains user ID, nickname, number of followers and followees, gender, birthday, education background, sexual orientation, marital status, blood type, zodiac, etc.

Additionally, we crawled $81,781,544$ location check-ins of these users in the two cities through Sina Weibo API[6]. Check-in records include check-in time, identity of the checked-in POI, and information of the POI (e.g. POI name, latitude, longitude, province, city and category, where the category follows a taxonomy [7]).

We note that all data we collected (including profile attributes and check-ins) is either through Open API or publicly available on Sina Weibo, i.e., no private data is used in the experiment.

2.2 Ground Truth Construction

To build a standard ground truth for our dataset, we are faced with several challenges. For one, out of privacy and safety concerns, some people tend to provide incomplete or even misleading profile information like gender and birthday. For another, nowadays, online social networks have many latent commercial zombie users with fake profiles. In order to address the above problems, we only consider verified accounts, since Sina corporation conducts manual verifications of these users (users are asked to submit officially authenticated materials such as copies of their ID cards and employment certificates) so as to make sure that these accounts provide real and authentic profile information.

Two matters should be noted: 1) verified users consist of a portion of celebrities, e.g. famous singer Taylor Swift and tech entrepreneur Bill Gates. 2) Sina corporation cannot inspect two subjective attributes, i.e. sexual orientation and marital status. For the former issue, we implement an auxiliary filter on the measurement of the number of followers of each user. Based on our finding, this approach effectively excludes celebrities, making our dataset more representative. For the latter issue, it is convinced that most normal users have the freedom not to offer these subjective attributes, but actually have no incentive to provide fake information.

As a result, we obtain $159,530$ verified users ($117,413$ Beijing users and $42,117$ Shanghai users) with both their profiles and check-in histories. Compared to many existing manually labeled experimental datasets [20, 23, 16], our approach generates a larger scale and more reliable dataset.

2.3 Demographics Description

Table 1 presents the completion rate (ratio of effective users) as well as the categories of different demographic attributes for the collected users. As is shown, most users provide their gender information, and a large portion of the users provide age, educational background and zodiac sign. Here, the raw educational attribute is provided as the name of universities or schools. For the sake

[6] http://bit.ly/1rgzRch
[7] http://bit.ly/1nnFztU

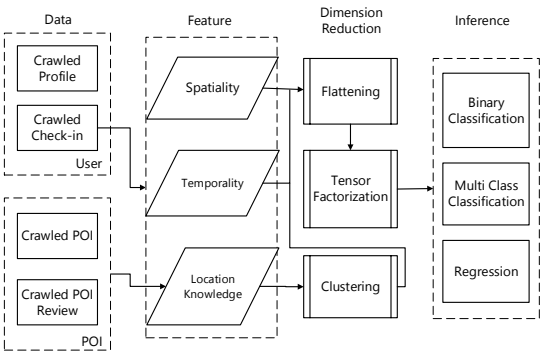
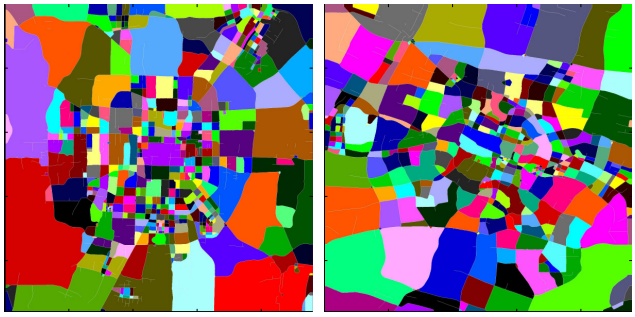


Figure 4: The Location to Profile (*L2P*) framework



(a) Beijing (b) Shanghai

Figure 5: Region segmentation for spatiality

of simplicity, we categorize it into two types: university and non-university for users who provide educational attributes. For marital status, we consider four statuses that have a dominant number of users in our dataset, while the others like divorced, widowed, separated, engaged etc. are excluded. Within the users who provide gender attributes, 45.81% are males, and 54.19% are females. Table 2 offers more statistics with respect to other attributes such as marital status, blood type, and sexual orientation.

3. LOCATION TO PROFILE *(L2P)*

In this section, we describe the *location to profile (L2P)* framework, which aims to predict a given users' demographic attributes based on his/her location check-ins. As illustrated in Figure 4, our proposed framework consists of four modules:

- The *data crawling module* accumulates user profiles and location check-ins with corresponding POI information on Sina Weibo and crawl customer reviews from Dianping.

- The *feature extraction module* extracts rich features from location check-ins of a given user in terms of spatiality, temporality and location knowledge, where the location knowledge is enriched with signals from both online review sites and social media, as detailed later in Section 3.1 and Section 3.2.

- The *dimension reduction module* incorporates the extracted features of users with a tensor model and performs tensor factorization to reduce the dimensionality of features, as demonstrated in Section 3.3.1.

- The *inference module* reduces the profile inference task to various classification and regression tasks and provides the final inference result, as detailed in Section 3.3.2.

3.1 Spatiality and Temporality

3.1.1 Spatiality

As indicated in Section 1, location check-ins are not uniformly distributed in the geo-spatial space. For a common person, his/her range of movement is mostly concentrated on several hubs like home, workplace, relatives' homes, etc. In addition, some businesses tend to be located in some concentrated areas, e.g., financial districts, high-tech districts and restaurant streets. Hence, people who frequently appear in these areas may exhibit similar demographic attributes.

To capture the spatial distributions of users' check-ins, we segment a city into disjointed regions. Each check-in of a user is assigned to the region that the check-in occurs in. Nevertheless, instead of using a uniform segmentation (grids) for a city, we adopt a morphological segmentation of urban spaces [31], where the regions are segmented using high level roads in a road network. Since transportation in urban areas is usually restricted by networks, such segmentation preserves the semantics of users' movements and the topology of road networks. Figures 5(a) and Figure 5(b) visualize the segmentation results of Beijing and Shanghai, where the segmented regions are indicated with different colors.

3.1.2 Temporality

Human location check-ins change over time during a day and a week. The underlying principle is that most people follow regular, predictable and stable patterns during their everyday lives, in spite of leading a host of different kinds of lifestyles.

Weekday and weekend patterns cover most people and nearly all kinds of jobs. Here, we split a week into two parts: weekdays and weekends. For both of them, we split a day into hourly time bins. Thus, we have a total of 24×2 time bins for the expression of temporal patterns. Similar to spatiality, we discretize the timestamps and assign the corresponding time bins for each check-in of a user.

3.2 Location Knowledge Enrichment

Here, we converge on the methodology of enriching the knowledge of a user's check-in location. Originally, a checked-in POI is associated with some basic information, e.g., the category. However, in order to predict user profiles, we need to extract more semantic and latent information for a POI. As a result, the enriched location knowledge of a given POI can be represented by a set of semantic features, such as the followings:

- Category feature. This feature shows the types and characteristics of location check-ins. All POIs follow a hierarchical taxonomy defined by Sina Weibo, and each POI is associated with a code, which is a 1-1 mapping to a certain category in the taxonomy.

- Review feature. Delivering users' interests and opinions on a POI, this feature is represented by scores of multi-aspects, e.g., when evaluating a restaurant, users are usually concerned with the atmosphere, service, taste and overall rating. We note that for different categories of POIs, the concerned aspects may vary.

- Keyword feature. Manifesting users' viewpoints toward a POI, this feature is a distribution over characteristic keywords (such as luxury, vintage, etc.) of a check-in location.

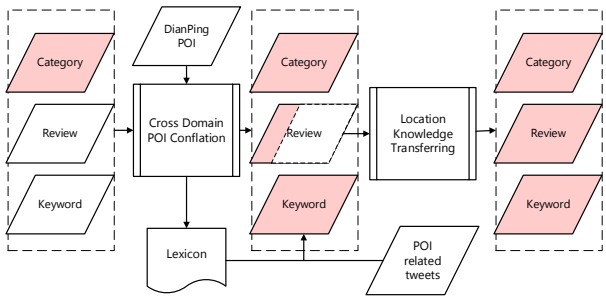


Figure 6: Location Knowledge Enrichment

Figure 6 sketches the location knowledge enrichment process. For the category feature, it is automatically attached to POI information in the crawled location check-in dataset. For learning the review and keyword features, we devise an incremental enrichment process. As is shown in Figure 6, the "location knowledge" of the POIs (indicated by the red part) grows as the process runs step by step. Specifically, we first conflate the POIs between Sina Weibo and Dianping, which is the largest user-generated review site in China. Later, we enrich the matched POIs with review features from Dianping. Next, we create a high utility keyword lexicon that contains relevant keywords for certain types of POIs (e.g., "delicious" is a high utility keyword for restaurants). Then, the keyword feature of all POIs can be derived from POI-related tweets on Sina Weibo based on the created lexicon. Finally, we train classifiers and regression models to transfer the review features from the matched POIs to the remaining unmatched POIs .

We note that in the above process, we utilize user generated data such as public reviews and related tweets of a POI, however, such side information is not directly related to an individual's check-ins, which are used as input for inferring users' profiles.

3.2.1 *Cross domain POI conflation*

Dianping and Sina Weibo are heterogeneous social networks facilitating users' different needs where Dianping is a specialized user review website while Sina Weibo is a social networking site. After crawling $2,699,822$ POIs with the corresponding $29,326,863$ reviews on Dianping, we devise a cross-domain POI conflation approach to link Dianping POIs with Sina Weibo POIs. Specifically, we extract the *name*, *address*, *telephone number*, *longitude* and *latitude* for POIs collected from both Dianping and Sina Weibo. As shown in Algorithm 1, for a pair of POIs $p_S^{(i)}$ (on Sina Weibo) and $p_D^{(j)}$ (on Dianping), we calculate Jaccard distance (1-Jaccard similarity) for name $d_N(p_S^{(i)}, p_D^{(j)})$ and address $d_A(p_S^{(i)}, p_D^{(j)})$, Hamming distance for telephone number $d_T(p_S^{(i)}, p_D^{(j)})$, and compute the geo-distance $d_G(p_S^{(i)}, p_D^{(j)})$ based on their latitudes and longitudes. As shown in Algorithm 1 lines 3-6, for each Sina Weibo POI $p_S^{(i)}$, we search for the POIs that are with the minimum distances in terms of different attributes (names, addresses, telephone numbers, geo-distances) from Dianping, and verify whether these POIs are matched to $p_S^{(i)}$ according to several fuzzy matching rules.

By applying the fuzzy matching algorithm, 35.26% POIs on Sina Weibo were matched to POIs on Dianping. In our practice, the thresholds in Algorithm 1 are chosen as follows: $\theta_A = \theta_N = 0.1$, $\theta_G = 50m$. To inspect the performance of the matching algorithm, we randomly sample 200 matched POIs, and manually check the correctness of the results, which achieve an accuracy of 99.5%.

Algorithm 1: Fuzzy POI Matching

Input: Sina Weibo POIs: $P_S = \{p_S^{(i)}\}_{i=1}^{q}$,
DianPing POIs: $P_D = \{p_D^{(j)}\}_{j=1}^{m}$
Output: $\{k_i\}_{i=1}^{q}$, where $p_S^{(i)}$ is matched to $p_D^{(k_i)}$ when $k_i > 0$, and $k_i = 0$ indicates no POI is matched to $p_S^{(i)}$

1 $J \leftarrow \{1, 2, \ldots, m\}$;
2 **for** $i = 1 : q$ **do**
3 $\quad j_T \leftarrow \arg\min_{j \in J} d_T(p_S^{(i)}, p_D^{(j)})$;
4 $\quad j_N \leftarrow \arg\min_{j \in J} d_N(p_S^{(i)}, p_D^{(j)})$;
5 $\quad j_A \leftarrow \arg\min_{j \in J} d_A(p_S^{(i)}, p_D^{(j)})$;
6 $\quad j_G \leftarrow \arg\min_{j \in J} d_G(p_S^{(i)}, p_D^{(j)})$;
7 $\quad$ **if** $d_T(p_S^{(i)}, p_D^{(j_T)}) = 0$ **then** /* identical phone number */
8 $\quad\quad k_i \leftarrow j_T, J \leftarrow J - \{j_T\}$;
9 $\quad$;
10 $\quad$ **else if** $d_G(p_S^{(i)}, p_D^{(j_N)}) \leq \theta_G$ **and** $d_A(p_S^{(i)}, p_D^{(j_N)}) \leq \theta_A$ **then**
$\quad\quad$ /* most similar name, close, and similar address */
11 $\quad\quad k_i \leftarrow j_N, J \leftarrow J - \{j_N\}$;
12 $\quad$ **else if** $d_N(p_S^{(i)}, p_D^{(j_G)}) \leq \theta_N$ **and** $d_A(p_S^{(i)}, p_D^{(j_G)}) \leq \theta_A$ **then**
$\quad\quad$ /* most close, similar name and address */
13 $\quad\quad k_i \leftarrow j_G, J \leftarrow J - \{j_G\}$;
14 $\quad$ **else if** $d_N(p_S^{(i)}, p_D^{(j_A)}) \leq \theta_N$ **and** $d_G(p_S^{(i)}, p_D^{(j_A)}) \leq \theta_G$ **then**
$\quad\quad$ /* most similar address, similar name and close */
15 $\quad\quad k_i \leftarrow j_A, J \leftarrow J - \{j_A\}$;
16 $\quad$ **else**
17 $\quad\quad k_i \leftarrow 0$;
18 **return** $\{k_i\}_{i=1}^{q}$

3.2.2 *Lexicon creation*

After the conflation of Dianping and Weibo POI, part of the Sina Weibo POIs are associated with corresponding reviews and users' aggregated scores for multi-aspects (like environment, taste, price etc.). Thus the review features for the matched POIs are already attained (as shown in Figure 6 where the review feature in the second phase is partially filled with red). The goal of this step is to create a high utility lexicon from users' review contents on Dianping. Specifically, we conduct Chinese word segmentation to segment each review to a set of words and eliminate the stop words. Since high utility keywords for different categories of POIs may vary (e.g., "delicious" is often used when talking about food in a restaurant, while "king-size" is usually used when talking about bed size in a hotel room), we categorize all POIs into several divisions as follows: *travel & accommodation*, *buildings & institutions*, *campus life*, *restaurant & delicacy*, *shopping & service*, *life & entertainment*, *park & outdoors*, *company* and *others*. Then for each division, we choose the top-n important words as the keywords (n is set to 200 in our implementation). Here, the importance of a term is defined by term frequency with a logarithmic factor which is widely used in information retrieval, given by:

$$tf_k = 1 + \log(f_k), \tag{1}$$

where f_k is the raw frequency of a keyword k in the reviews of a certain POI. As a result, we obtain a set of keywords for each category division, and the lexicon is merged by the keywords of all sets.

Meanwhile, we collect POI-related tweets for all crawled POIs on Sina Weibo using the open API [8]. Next, for each POI (matched or unmatched), we generate the keyword features by computing the distribution over the keywords in the created lexicon using the POI-related tweets (pre-processed with Chinese word segmentation and

[8] http://bit.ly/1rQQZXe

stop words filtering). Hence, we obtain the keyword features for all POIs.

3.2.3 Location knowledge transferring

As indicated in Section 3.2.1, only a fraction (35.26%) of Weibo POIs have been matched, not to mention that actually some unpopular matching POIs have vacant review scores for different aspects. Since users' review keywords fully embody their ratings, transferring the knowledge of review keywords as ratings is intuitive. In order to obtain the review features for all collected POIs, we develop a supervised learning approach to "transfer" the review features of matched POIs that have multi-aspect review scores to the remaining POIs.

Specifically, given the keyword feature of a POI F_K, the score of a certain aspect is considered a function of the keyword features, i.e., $r_i = r_i(F_K)$, where $i = 1, 2, \ldots, M$ is the index of all review aspects (merged from different category divisions) and $r_{all} = r_{all}(F_K)$ is the overall rating of a POI. The key insight behind this is that the multi-aspect scores provided by users are typically reflected from their reviews. Thus, we utilize the matched POIs that have multi-aspect review scores as the training data to learn the mapping functions $\{r_i\}_{i=1}^{M}$ as well as the overall rating r_{all}, where the review score is deemed as the label and the keyword feature of a POI is considered as the input feature.

We note that the score for a single review aspect is a categorical variable $(1, 2, \ldots, 5)$ indicated by "stars" and the overall rating of a POI is a continuous valued rating. Thus, we train linear kernel SVM classifiers for individual review aspects and logistic regression to learn the overall rating of a POI. We conduct a 10-fold cross-validation on the POIs using the trained classiers and regression models. Then, we apply the trained models to infer the aspect scores and overall ratings for the remaining POIs that have no review features. As a result, for each POI, we obtain the review feature $F_R = (r_1, r_2, \ldots, r_M, r_{all})$.

3.3 Dimension Reduction and Profile Inference

3.3.1 Dimension reduction

After feature extraction from *spatiality*, *temporality* and *location knowledge* aspects, we obtain a large scale feature set for users' check-ins. To reduce the dimensionality of the features and mine intrinsic representation of users, we feed the features into a three-way tensor T and conduct tensor factorization.

As shown in Figure 7, the three-ways of $T \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ represent users, location knowledge features and contextual features respectively, where I_1 is the number of users, I_2 is the dimensionality of location knowledge features and I_3 is the dimensionality of the contextual features including both spatiality and temporality.

Here, in order to construct the three-way tensor, Cartesian production is implemented over feature spaces within context dimension and location knowledge dimension severally. More specifically, for context dimension, given a check-in corresponding to region α and time bin β, the index is determined by $\alpha \times N_T + \beta$, in which N_T is the number of time bins. For the dimension of location knowledge, we first cluster check-ins according to review features and keyword features separately (we employ k-means clustering in our practice while other clustering methods can generally be applied), resulting in N_R review-based clusters and N_K keyword-based clusters respectively. Then we flatten category, review, and keyword features using an approach similar to the context dimension. For example, assuming a user's check-in c is associated with a POI that belongs to category α_1, cluster α_2 of the N_R review-based clusters, and cluster α_3 of the N_K keyword-based clusters respectively, the final coordinate of c in the location knowledge dimension is given by $\alpha_1 \times N_R \times N_K + \alpha_2 \times N_K + \alpha_3$.

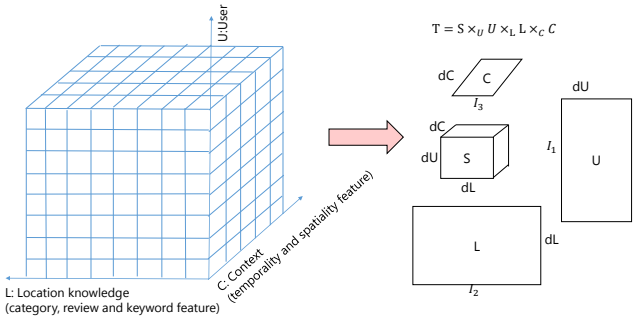


Figure 7: Tensor decomposition model

Later, we employ the Tucker Decomposition technique which decomposes T by:

$$T = S \times_U \mathrm{U} \times_L \mathrm{L} \times_C \mathrm{C},$$

where $S \in \mathbb{R}^{dU \times dL \times dC}$ is the core tensor indicating interactions between user, context and location knowledge. $U \in \mathbb{R}^{I_1 \times dU}$, $L \in \mathbb{R}^{I_2 \times dL}$ and $C \in \mathbb{R}^{I_3 \times dC}$ are the factor matrices considering the low dimensional representation of users, locations and contexts (spatiality and temporality) respectively.

The tensor model offers several advantages for dimension reduction compared to other approaches, e.g., the tensor model intrinsically captures interactions between different features within various domains similar to matrix factorization. However, compared to matrix factorization, the tensor model is more suitable for high dimensional data.

3.3.2 Profile inference

Let U be the factorized user intrinsic matrix generated using the method described in Section 3.3.1, which already incorporates users' similar patterns in terms of spatiality, temporality, and location knowledge. By considering each row of U as the feature vector of a user and the actual demographic attribute as the label, we transform the profile inference task to several classification or regression tasks as follows.

- Gender. A binary classification task classifying users' gender: *male* and *female*.
- Age: A regression task inferring the specific age, where age is a continuous valued attribute.
- Education Background. A binary classification task distinguishing users' education level: *university* and *non-university*.
- Sexual Orientation. A multi-class classification task differentiating four kinds of sexual orientation: *heterosexuality*, *bisexuality*, *male homosexuality* and *female homosexuality*.
- Marital Status. A multi-class classification task predicting four kinds of marital status: *single*, *courtship*, *in love* and *married*.
- Blood Type. A multi-class classification task identifying four common blood types: *O*, *A*, *B* and *AB*.
- Zodiac: A multi-class classification task demarcating twelve zodiac signs.

Over the last several decades, various algorithms have been proposed for learning classification and regression functions. We evaluate a number of well-adopted learning algorithms to learn the models separately and examine their performance in the experiments. Specifically, for binary classification tasks, we compare logistic regression, SVM, and LambdaMART [28]; for regression tasks,

we try models such as linear regression, Poisson regression, and boosted tree regression [9]; for multi-class classification, we evaluate multi-class logistic regression, multi-class neuron networks [19], and a parallel ensemble method [24] (refer to Section 4 for the results).

4. EXPERIMENTS

4.1 Settings

4.1.1 Preliminary

Our experiments are guided by the following considerations: 1) Whether location check-ins have predictive power for demographic attributes; 2) Whether all the proposed features (spatiality, temporality and location knowledge) contribute to the performance of inferring demographic attributes; and 3) How well the proposed *L2P* framework applies to different demographic attributes.

With the purpose of testing the generality and robustness of our L2P framwork, we perform experiments on Beijing and Shanghai separately. For each city, we conducted a 10-fold cross validation for inferring each demographic attribute. For the profile inference procedure (introduced in Section 3.3.2), we used 6 parts for training, 2 parts for validation, and the remaining 2 parts for testing if not specified elsewhere (we also study the effect of varying the ratio of training data in the experiments). The validation set was used to tune the hyper-parameters of the learning algorithms, such as the number of trees in LambdaMART, and then we fixed the parameters to train the models using the training set. The test set was used to evaluate the performance of the trained models. Note that all the reported results are the average performances over the 10 trials.

For the parameters, we set $n = 200$ for choosing the top-n keywords in lexicon creation (Section 3.2.2). In the procedure of location knowledge transferring (Section 3.2.3), N_R and N_K are set to 100 by default, and dU is set to 200 by default if not specified elsewhere in Section 4.2. The effect of different N_R, N_K, and dU are studied later in the experiments.

4.1.2 Baselines

To the best of our knowledge, there exists no model directly predicting user demographics from human location check-ins. We then compare our method with the following baselines. Here, by analogizing a POI as a "Like", we compare our approach against a recently famous profile inference method proposed by Kosinski et al. [13], which extracts principle components from the User-POI matrix and performs logistic/linear regression based on the derived user feature vector. We refer to this baseline as the POI-based method (shortened as **POI** method). We term our approach that comprehensively considers spatiality, temporality, and location knowledge as the **STL** method, and further evaluate the following baselines for comparison.

- Spatiality-based method (**S**), which only considers the spatial information in the tensor (degenerated as a matrix), and performs profile inference using extracted user features similar to the STL method.

- Spatiality and Temporality-based method (**ST**), which goes one step further than S, leveraging both the spatial and temporal features in the tensor for prediction.

- Spatiality, Temporality, and Category-based method (**STC**), which adopts spatial, temporal and category features in the tensor for prediction, i.e., without location knowledge enrichment as in the STL method.

4.1.3 Criteria

We evaluate the results using the following measurements in terms of different inference tasks. Effective and common metrics for binary classification such as precision, recall, F-measure and Area Under the ROC Curve (AUC) are all evaluated. For multi-class classification, average precision (over different classes), average recall, and average F-measure are employed. We use RMSE (Root-Mean-Square Error) for examining the performance of regression models.

4.2 Results

4.2.1 Gender

Table 3 presents the results of gender inference for users in Beijing and Shanghai. As is shown in Table 3(a), the STL method achieves the best performance in terms of all measurements, where the F1 score is above 0.8 for both Beijing and Shanghai users, which significantly outperforms competing methods. In particular, the STL method, which considers the complement set of location knowledge features, gains more than 0.05 improvement over the STC method and 0.1 improvement over the POI method in terms of AUC for both Beijing and Shanghai users. This indicates that the location semantics, particularly, the review and keyword features learned from different domains, contribute a lot to the predictive power of the STL method.

We further study the performance of different learning algorithms (Section 3.3.2) for gender inference, and the dimensions of user intrinsic matrix dU in tensor factorization (Section 3.3.1) for the STL method. Figure 8 illustrates the AUC for SVM (linear kernel), LR and LambdaMart classifiers. As is shown, the LambdaMART classifier performs the best over other classifiers. As the number of dimensions dU increases from 50 to 200, the results of all classifiers improve and tend to be stable when $dU = 300$, then decrease when dU further increases.

Table 3: Performance of Gender Inference

(a) Beijing

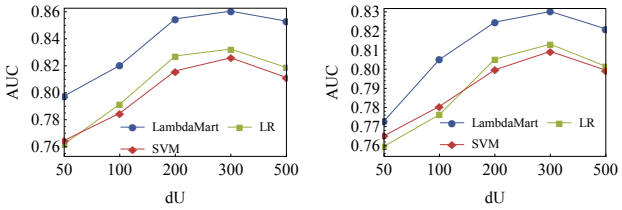
	Precision	Recall	F1	AUC
POI	0.7102	0.7055	0.7078	0.7502
S	0.6921	0.6899	0.6910	0.7321
ST	0.7321	0.7429	0.7375	0.7746
STC	0.7727	0.7631	0.7679	0.8027
STL	**0.8211**	**0.8059**	**0.8134**	**0.8548**

(b) Shanghai

	Precision	Recall	F1	AUC
POI	0.7362	0.7434	0.7398	0.7463
S	0.7197	0.7218	0.7207	0.7266
ST	0.7528	0.7596	0.7562	0.7682
STC	0.7819	0.7704	0.7761	0.8151
STL	**0.8368**	**0.8127**	**0.8246**	**0.8654**

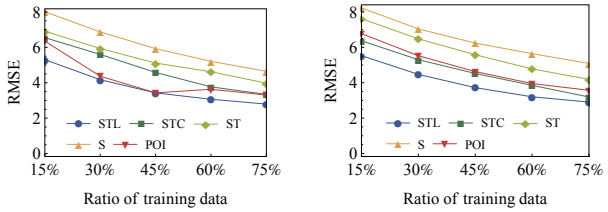
4.2.2 Age

Age inference is considered a regression task. Figure 9 plots the RMSE of all compared methods when changing the ratio of training data. As is shown, the STL method and POI method perform better than other methods for age inference, but STL still outperforms the POI method. For example, when we use only 15% data for training, the RMSE of the STL method is less than 5, which is a significant improvement over the POI method. Table 4 gives the results of different regression models for STL, where the boosted tree regression model achieves the best performance.



(a) Beijing (b) Shanghai

Figure 8: Performance of different classifiers for STL changing over *dU*



(a) Beijing (b) Shanghai

Figure 9: RMSE w.r.t. ratio of training data

Table 4: RMSE of Different Regression Models for STL

	Beijing	Shanghai
Poisson	3.463	3.602
Linear	3.312	3.399
Boosted Tree	**3.067**	**3.204**

4.2.3 Education background

The results of education background inference are presented in Table 5. Surprisingly, the inference results of all methods are remarkably higher than gender and age, which indicate that users' education levels are highly predictable from their physical movements. For example, the AUC of STL method for Beijing users is higher than 0.9. Meanwhile, we study the effect of N_K (number of keyword-based clusters) and N_R (number of review-based clusters) for dimension reduction in location knowledge enrichment. As shown in Table 6, the performance is comparatively stable for different settings of N_R and N_K.

4.2.4 Sexual orientation and marital status

Table 7 presents the evaluation results for inferring sexual orientation and marital status. Here, due to space limit, we report the overall results for Beijing and Shanghai users, and based on the best-performed parallel ensemble classifier. We note that for both sexual orientation and marital status, the inference problem is reduced as a 4-class classification task. Although the performance is not as strong as the gender/education inference, the STL method still outperforms other methods in terms of all measurements, and for all compared methods the results of sexual orientation inference are better than marital status.

4.2.5 Blood type and zodiac sign

We note that for blood type and zodiac inference, which have rarely been considered in existing literatures our work is rather initiatory and purely driven by curiosity. As is shown in Table 8, the location check-ins show weak predictive power for these attributes.

Remark: From the above inference results, we can discover that gender and education attributes achieve the best results because they are highly predictable from check-in records and the binary classification is relatively easy. Considering the identification of sexual orientation and marital status, since they are multi-class classification tasks and the attribute completion rate is very low, the prediction is not as good as gender and education. Careful scrutiny reaveals that sexual orientation inference is better than that of marital status, because sexual orientation (such as bisexuality and homosexuality) is more distinctive than marital status (such as single and courtship). The performances of blood type and zodiac sign are the weakest among all the demographics, indicating that they are merely predictable from human location check-in records. However, we should note that our model still perfoms much better than random guess. As is shown in the above prediction tasks, an ascending order of the predictive power for the compared methods are S, POI, ST, STC and STL. Spatial feature, which is extracted based on road networks in coarse granularity, leads to weaker performance than POI. On the basis of spatiality (S), we iteratively add tempo-

Table 5: Performance of Education Background Inference

(a) Beijing

	Precision	Recall	F1	AUC
POI	0.7564	0.7702	0.7632	0.7992
S	0.7385	0.7294	0.7339	0.7723
ST	0.7655	0.7702	0.7678	0.8150
STC	0.8073	0.7921	0.7996	0.8413
STL	**0.8774**	**0.8829**	**0.8801**	**0.9021**

(b) Shanghai

	Precision	Recall	F1	AUC
POI	0.7759	0.7817	0.7788	0.8205
S	0.7394	0.7191	0.7291	0.7659
ST	0.7804	0.7631	0.7717	0.8041
STC	0.8115	0.8058	0.8086	0.8503
STL	**0.8823**	**0.8726**	**0.8774**	**0.8957**

Table 6: Performance w.r.t. Number of Clusters in Dimension Reduction

	Beijing		Shanghai	
	F1	AUC	F1	AUC
$N_R = 100, N_K = 100$	0.8801	0.9021	0.8774	0.8957
$N_R = 100, N_K = 200$	0.8812	**0.9100**	0.8762	0.8946
$N_R = 200, N_K = 100$	0.8792	0.8977	**0.8801**	0.8896
$N_R = 200, N_K = 200$	**0.8867**	0.9017	0.8792	**0.9051**

Table 7: Sexual Orientation and Marital Status Inference

	Sexual Orientation			Marital Status		
	Precision	Recall	F1	Precision	Recall	F1
POI	0.4637	0.4586	0.4611	0.3257	0.3302	0.3279
S	0.4496	0.4503	0.4499	0.3291	0.3287	0.3289
ST	0.4613	0.4599	0.4606	0.3451	0.3502	0.3476
STC	0.4826	0.4855	0.4840	0.3683	0.3677	0.3680
STL	**0.5235**	**0.5122**	**0.5178**	**0.3855**	**0.3789**	**0.3822**

Table 8: Blood Type and Zodiac Inference

	Blood Type			Zodiac		
	Precision	Recall	F1	Precision	Recall	F1
POI	0.2801	0.2729	0.2765	0.1203	0.1236	0.1219
S	0.2744	0.2738	0.2741	0.1103	0.1052	0.1077
ST	0.2775	0.2742	0.2758	0.1256	0.1260	0.1258
STC	0.2894	0.2901	0.2897	0.1286	0.1243	0.1264
STL	**0.3012**	**0.3103**	**0.3057**	**0.1303**	**0.1275**	**0.1289**

raility (ST), category (STC), and location knowledge (STL), making the feature more sophisticated step by step. ST and STC baselines, which convey temporal patterns, result in enhanced achievements over spatiality-based methods (S and POI), where the STC method partially considers location knowledge from the category information of a POI. The STL method, which incorporates more location knowledge transfered from review sites, gains the most remarkable improvement against all other baselines.

5. RELATED WORK

5.1 Inferences of Demographic Attributes

Demographics inference has been studied in academia for more than fifty years. Several early stage works have researched author identification [17] and gender discrimination [14, 26]. In 1992, Costa Jr and McCrae [8] predicted demographic attributes and personalities through people's answers to a specific psychometric test. [11] adopted psychological approach and analyzed testers' written words to classify their personality type.

The proliferation of digital communication and Internet brought a brand new opportunity for inferring demographic attributes. With a concentration on differences in gender sociolinguistics, Boulis and Ostendorf [5] studied users' telephone conversation data and predicted gender attributes using machine learning techniques. Murray and Durrell [18] adopted LSA to analyze users' browsing behavior for demographic attributes classification. Back et al. [3] predicted personality, i.e. neuroticism, openness, agreeableness, conscientiousness, narcissism, and extroversion solely from email addresses.

In recent years, the fast development of online social networks and mobile computing technologies have provided researchers with a convenient way to infer people's demographic attributes. Kosinski et al. [13] employed Facebook "Likes" data, SVD dimension reduction technique and logistic regression to predict private traits and attributes. Pennacchiotti and Popescu [20] put forward a machine learning framework to detect users' political affiliation, ethnicity identification, and business affinity. Brdar et al. [6] proposed a multi-level classification model to predict people's attributes like gender, age, marital status, and job type, among others.

Our work differs from existing approaches in that we focus on users' location check-in data from online social networks for demographic inference, which has rarely been explored before.

5.2 Location Understanding

Existing approaches on enriching location knowledge can be categorized in terms of different data sources: *GPS trace*, *Web* and *location-based social networks (LBSN), and Volunteered Geographic Information (VGI)*.

• GPS trace: A myriad of scholars have carefully researched into location knowledge extraction from GPS traces [7, 15]. They employed and devised different methods for mining and extracting semantically important locations.

• Web: An early location semantics study [21] addressed the deficiency of traditional physical/geographical location presentation by linking location to an unambiguous "web-like" URI. Focusing on location extraction from web contextual information, Qin et al. [22] proposed a detection ranking framework and addressed challenges in location name detection and location entity disambiguation. Wang et al. [27] proposed a generative model to learn latent aspect ratings from online reviews, demonstrating the effectiveness of the proposed model using data from a hotel review site.

• LBSN and VGI: Recently, several works have been proposed on location knowledge mining in LBSN and VGI. Ye et al. [30] developed techniques for location semantic annotation by extracting features of places from *explicit patterns (EP)* and *implicit relatedness* and learning SVM to support multi-label classification. [21] and [29] respectively presented methodologies to analyze spatial and temporal dimension semantics of POI from LBSN and VGI.

Our location knowledge enrichment is distinctive from existing works in the following ways: 1) We enrich the understanding of a POI by transferring location knowledge from heterogeneous domains, after conflating POIs between an online customer review site and an online social network; 2) We explicitly model three types of location knowledge, namely category features, review features, and keyword features, where review features and keyword features are learned from heterogeneous user generated data (online reviews and POI-related tweets).

5.3 Tensor factorization

Tensor factorization is widely applied in a variety of areas. The first application in data mining is [1] in 2005, focusing on chat room tensor construction and factorization techniques performance comparisons. Later in text analysis, Bader et al. [4] developed nonnegative tensor factorization to extract discussions from email communications and proved the superiority of nonnegative matrix factorization. The application of tensor factorization on recommendation and prediction is also carefully studied in several works [2, 25] where Acar et al. [2] explored matrix and tensor factorization models in temporal link prediction and Symeonidis et al. [25] proposed a unified framework modeling 3-order tensor, user, item and tag.

To the best of our knowledge, our work is the first to apply tensor factorization for profile inference in real world online social networks where features are organized in a three-way tensor, which consists of user, context (spatiality and temporality) and location knowledge.

6. CONCLUSION AND DISCUSSIONS

In this paper, we have addressed the problem of demographics inference from location check-ins, and developed a comprehensive *L2P* framework to capture temporality, spatiality and location knowledge at the same time. Extensive experiments based on our large-scale Sina Weibo dataset have validated the effectiveness of the proposed methods for inferring a variety of demographic attributes.

Although this is, to our best knowledge, the first work that explores demographic inference from location check-ins, the proposed methodology is far from perfect and may suffer from several limitations. For example, our experiments are based on a dataset with users who use online social networks and post check-ins, which may not be a representative sample set of the whole targeted population. Another issue is the profiles of users who live in metropolitan cities (Beijing and Shanghai in our dataset) may create bias in the inference result. However, we believe that the proposed *L2P* framework is general enough to be applied on other forms of human mobility data with a finer granularity and larger coverage, e.g., GPS traces collected from phones.

There are many directions that can be further explored to improve the current *L2P* framework. For one thing, location check-ins may suffer from data sparsity for representing human mobility, given that few people are willing to post locations all the time. One possible solution may come from extracting POI-level locations from a large quantity of normal tweets, like "I am shopping with my daughter at Walmart in L.A.!". It is also possible that other signals could be integrated into the *L2P* framework, e.g., the current model has not exploited users' friendship relations on social networks, which might affect users' check-in behaviors.

References

[1] E. Acar, S. A. Camtepe, M. S. Krishnamoorthy, and B. Yener. Modeling and multiway analysis of chatroom tensors. In *Intelligence and Security Informatics*, pages 256–268. Springer, 2005.

[2] E. Acar, D. M. Dunlavy, and T. G. Kolda. Link prediction on evolving data using matrix and tensor factorizations. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pages 262–269. IEEE, 2009.

[3] M. D. Back, S. C. Schmukle, and B. Egloff. How extraverted is honey. bunny77@ hotmail. de? inferring personality from e-mail addresses. *Journal of Research in Personality*, 42(4): 1116–1122, 2008.

[4] B. W. Bader, M. W. Berry, and M. Browne. Discussion tracking in enron email using parafac. In *Survey of Text Mining II*, pages 147–163. Springer, 2008.

[5] C. Boulis and M. Ostendorf. A quantitative analysis of lexical differences between genders in telephone conversations. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 435–442. Association for Computational Linguistics, 2005.

[6] S. Brdar, D. Culibrk, and V. Crnojevic. Demographic attributes prediction on the real-world mobile data. In *Proc. Mobile Data Challenge by Nokia Workshop, in Conjunction with Int. Conf. on Pervasive Computing, Newcastle, UK*, 2012.

[7] X. Cao, G. Cong, and C. S. Jensen. Mining significant semantic locations from gps data. *Proceedings of the VLDB Endowment*, 3(1-2):1009–1020, 2010.

[8] P. T. Costa Jr and R. R. McCrae. Reply to ben-porath and waller. *Psychological Assessment*, 4(1):20–22, 1992.

[9] G. De'Ath. Boosted trees for ecological modeling and prediction. *Ecology*, 88(1):243–251, 2007.

[10] Y. Dong, Y. Yang, J. Tang, and N. V. Chawla. Inferring user demographics and social strategies in mobile social networks. 2014.

[11] L. A. Fast and D. C. Funder. Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. *Journal of personality and social psychology*, 94 (2):334, 2008.

[12] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[13] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[14] W. Labov. *The social stratification of English in New York City.* PhD thesis, Columbia university., 1964.

[15] J. Liu, O. Wolfson, and H. Yin. Extracting semantic location from outdoor positioning systems. In *MDM*, page 73, 2006.

[16] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260. ACM, 2010.

[17] F. Mosteller and D. L. Wallace. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302): 275–309, 1963.

[18] D. Murray and K. Durrell. Inferring demographic attributes of anonymous internet users. In *Web Usage Analysis and User Profiling*, pages 7–20. Springer, 2000.

[19] G. Ou and Y. L. Murphey. Multi-class pattern classification using neural networks. *Pattern Recognition*, 40(1):4–18, 2007.

[20] 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*, pages 430–438. ACM, 2011.

[21] S. Pradhan. Semantic location. *Personal Technologies*, 4(4): 213–216, 2000.

[22] T. Qin, R. Xiao, L. Fang, X. Xie, and L. Zhang. An efficient location extraction algorithm by leveraging web contextual information. In *proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 53–60. ACM, 2010.

[23] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010.

[24] L. Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.

[25] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. A unified framework for providing recommendations in social tagging systems based on ternary semantic analysis. *Knowledge and Data Engineering, IEEE Transactions on*, 22 (2):179–192, 2010.

[26] P. Trudgill. *The social differentiation of English in Norwich*, volume 13. CUP Archive, 1974.

[27] H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 618–626. ACM, 2011.

[28] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270, 2010.

[29] M. Ye, K. Janowicz, C. Mülligann, and W.-C. Lee. What you are is when you are: the temporal dimension of feature types in location-based social networks. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 102–111. ACM, 2011.

[30] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz. On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 520–528. ACM, 2011.

[31] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 186–194. ACM, 2012.

[32] N. J. Yuan, F. Zhang, D. Lian, K. Zheng, S. Yu, and X. Xie. We know how you live: exploring the spectrum of urban lifestyles. In *Proceedings of the first ACM conference on Online social networks*, pages 3–14. ACM, 2013.

[33] E. Zheleva and L. Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th international conference on World wide web*, pages 531–540. ACM, 2009.