2016 IEEE 16th International Conference on Data Mining

HLGPS: A Home Location Global Positioning System in Location-Based Social Networks

Yulong Gu, Jiaxing Song, Weidong Liu†, Lixin Zou
Department of Computer Science and Technology
Tsinghua University, Beijing, 100084, China
guyulongcs@gmail.com, {jxsong, liuwd}@tsinghua.edu.cn, zoulx15@mails.tsinghua.edu.cn

***Abstract*—The rapid spread of mobile internet and location-acquisition technologies have led to the increasing popularity of Location-Based Social Networks(LBSNs). Users in LBSNs can share their life by checking in at various venues at any time. In LBSNs, identifying home locations of users is significant for effective location-based services like personalized search, targeted advertisement, local recommendation and so on. In this paper, we propose a Home Location Global Positioning System called HLGPS to tackle with the home location identification problem in LBSNs. Firstly, HLGPS uses an influence model named as IME to model edges in LBSNs. Then HLGPS uses a global iteration algorithm based on IME model to position home location of users so that the joint probability of generating all the edges in LBSNs is maximum. Extensive experiments on a large real-world LBSN dataset demonstrate that HLGPS significantly outperforms state-of-the-art methods by 14.7%.**

***Keywords*—Home Location Identification, Location-Based Social Networks, Influence Model, Social Networks**

I. Introduction

In recent years, we have seen a rapid proliferation of Social Networks such as Facebook, Twitter, Google+ and so on. The rapid growth of mobile internet and location-acquisition technologies have led to the increasing popularity of Location-Based Social Networks(LBSNs) such as Foursquare[1], Facebook Places[2] and so on by embedding location into Social Networks. Users in LBSNs can conveniently log their activity histories with spatio-temporal data by checking in at various venues(e.g., restaurants, airports, scenic spots) at any time using their smart phones. As the rapid growth of LBSNs, Home Location Identification of users becomes one of the most important problems because home location of users are extremely important for various applications to provide effective location-based services. For example, profiling users' home locations enables search engines to provide personalized search results in mother tongue of users, news sites to recommend localized news and advertisers to recommend local advertisements.

The home location of a user is defined as the "permanent" place where the user spend most of his time in as previous work[9]. It captures the major and static geographic scope of the user and therefore provides valuable information for personalized services. However, the home location problem is quite challenging. Firstly, only a small percentage of users provide their home locations in Social Networks due to privacy concerns. On twitter, only a few people (16%) register city level locations in their profiles and most of users leave general, non-sensical or even blank information[9]. Secondly, users may check in at various places far from their home and make friends far away. Thirdly, many users do not have any check-in data. As of September 2013, only 30% of users provide their location information to at least one social media account and 12% of adult smartphone owners have used geo-social services to check in at some location[3]. This problem has been attracting great interests of researchers in academic recently[2, 3, 8–10, 15]. Existing approaches can mainly be divided into two parts: Content based approach [2, 3, 8] and Check-in based approach[9, 10, 15]. Content based approach infers home location of users using models based on extracted location information from texts like tweets in Social Networks. Check-in based approach infers home location of users leveraging check-in data of users. However, these methods are still not effective enough.

In this paper, we propose a Home Location Global Positioning System called HLGPS to tackle with the tough home location identification problem in LBSNs. HLGPS uses an influence model named as IME to model the edges in LBSNs. Specifically, IME is an unified probabilistic model that models edges in LBSNs based on signals from Social relationship data(social friendship and social trust) and User-centric data(check-in data and rating data). We represent a LBSN as a directed heterogenous graph where the nodes can be users or venues. Edges in the graph can be following edges between users, check-in or rating edges from users to venues. In IME, we model each node with a location and an influence scope. We assume each edge $t \rightarrow h$ from a tail node t to a head node h is generated according to locations of h and t, influence scope of the head node h, social trust value of the head node h for the tail node t. IME is based on the ideas that people tend to make friends with people living near or people who have more common friends with them, follow celebrities, visit popular places, check in at venues nearby and rate venues

†Corresponding author: Weidong Liu (E-mail: liuwd@tsinghua.edu.cn)

[1]https://foursquare.com
[2]https://www.facebook.com/places/
[3]http://www.pewinternet.org/2013/09/12/location-based-services

2374-8486/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDM.2016.144

that are near to them. In this paper, we propose the idea of using “social trust” to measure closeness in social structure to model edges in LBSNs. Social trust between users is measured by applying sigmoid function on the number of their common friends. It will be higher if two users have more common friends. People will tend to rate venues if they visit them often and be familiar with them. These venues will probably be near to their home. To the best of our knowledge, we are the first that propose the idea of using “social trust" and “rating data" for Home Location Identification problem. HLGPS then uses a two-stages global iteration method to position home location of users based on IME model. In the first stage, for users who have check-in data, we develop a single-pass clustering algorithm to cluster their check-in data and select the center of largest cluster as home locations of them. In the second stage, we use a global iteration algorithm to estimate home location of users so that the joint probability of generating all the edges in LBSNs is maximum.

We have conducted extensive experiments to evaluate HLGPS and compared it with state-of-the-art methods[4, 9, 10, 15, 17] based on a large-scale Foursquare dataset containing about 836K users and 649K venues. Experiment results show that HLGPS can predict home locations of users who have check-in data at the accuracy 92.3% though the average check-in number of each user is only about 2.7. HLGPS can predict home locations of all users who don’t have home location at the accuracy of 67.7%, outperform state-of-the-art methods by about 14.7%, when only 16.7% users have check-in data. In a word, out method significantly outperforms state-of-the-art methods, and achieve the best performance.

II. Dataset Description

A. Foursquare Dataset

Foursquare is currently one of the largest and most popular LBSNs. Users in Foursquare can share their locations with friends and followers through check ins. As of December 2013, Foursquare had 45 million registered users[4]. In this paper, we use a widely used and publicly available Foursquare dataset [6, 16]. In this dataset, each user has an unique id and a geospatial location that represents the user’s home town location. Each venue has an unique id and a geospatial location. The social graph data contains the social graph edges that exist between users. The rating data consists of implicit ratings that quantify how much a user likes a specific venue. In the Foursquare dataset within the continental United States, there are 835,896 users, 648,825 venues, 370,477 check-ins, 1,397,412 ratings and 12,924,609 social graph edges.

B. Mapping Location to City

In this paper, we map a location to a specific city in following method: The candidate cities which we select are the 297 cities in the United States with a population of at least 100,000 on July 1, 2014, as estimated by the United States Census Bureau[5]. We define a location’s mapped city as the nearest candidate city from the location.

[4]https://en.wikipedia.org/wiki/ Foursquare

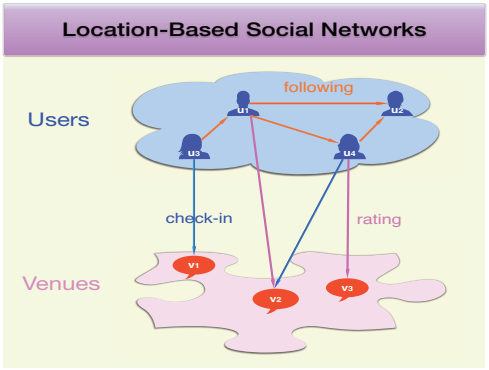

Fig. 1. Location-Based Social Networks(LBSN)

III. Home Location Identification Problem Formulation

A. Location-based Social Networks Formulation

We represent a Location-Based Social Network as a directed heterogeneous graph $G = (N, E)$. Demonstration of the LBSN is shown in Figure 1. We denote nodes and edges in the graph as N, E. For the nodes, we denote users and venues as U, V. We use U^H and U^{-H} to denote the set of users whose home locations are known and not known. Similarly, we use U^C and U^{-C} to denote the set of users who have and not have check-in data. For a geographical location L denoted by (α, β), α is the latitude and β is the longitude. We denote the home location of user u_i as L_{u_i} and location of venue v_j as L_{v_j}. The geographical distance between nodes n_i and n_j is denoted as $dis(n_i, n_j)$. For the edges, we denote following edges, check-in edges, ratings as F, C, R. Further, we denote incoming and outgoing nodes of node n of edge type t as $I_t(n)$ and $O_t(n)$.

B. Home Location Identification Problem Formulation

Home Location Identification Problem For a Location-based Social Network $G = (N, E)$, for each user in U^{-H}, estimate a home location $\widetilde{L}_{u_i}$ so as to make $\widetilde{L}_{u_i}$ close to u_i’s true home location L_{u_i}. The problem is formatted as Equation 1.

$$\min_{\widetilde{L}_u} \frac{1}{|U^{-H}|} \sum_{u_i \in U^{-H}} dis(L_{u_i}, \widetilde{L}_{u_i}) \tag{1}$$

IV. IME: Influence Model on Edges

Our Home Location Global Positioning System HLGPS uses an influence model names as IME to model edges in Location-Based Social Networks.

[5]https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population

A. Motivation of IME model

In this paper, we exploit social friendship, check-in data, social trust and rating data to model edges in social networks.

1) social friendship: The probability of friendship decreases as the distance between nodes increases has been observed from social networks like Facebook, Twitter and so on[1, 9]. Different nodes have different influence and they have different probabilities to attract tail nodes at the same distance[9]. For example, a star is more likely to attract users who live far away than a regular user.

2) check-in data: Users tend to check-in at venues near to them[4, 9]. [4] infer the home location by discretizing the world into cells and defining the home location as the average position of check-ins in the cell with the most check-ins. [14] take the most popular place where the user check in as her home location. Consequently, for users who have check-in data, we can directly predict home location of users leveraging their check-in data.

3) social trust: Existing methods[9] consider friend relation as a binary relationship. However, closer friends in social networks should have more influence on the home location of users. In this paper, we propose the concept "social trust" to measure the closeness in social structure and firstly apply it to Home Location Identification problem. We denote the social trust value of node n_i for node n_j as ST_{ji} and measure social trust between nodes by applying sigmoid function($sigmoid(x) = \frac{1}{1+e^{-x}}$) on the number of common friends. In the experiments, we also try the commonly used metric Jaccard Similarity[7] performing on friend set of users to measure social trust.

4) rating data: In this paper, we firstly introduce the using of "rating" data for Home Location Identification problem. If a user rate a venue, he may visit the place often or be familiar with the place. The venues is probably near to the user who rate them.

B. Formulation of IME Model

In the IME model, we denote the influence of a node n_i as I_{n_i} which is a probability distribution over the geographic plane. For a node n_i, we define n_i's influence on another node n_j at a location L as the probability that n_j build an edge $e\langle n_j, n_i\rangle$ to it. An influential node (user or venue) will have more broad influence scope and more influence at the same distance than an ordinary node.

1) Influence Model of Nodes on Geographic: We choose a gaussian distribution to capture a node's influence model for its expressiveness and simplicity [9]. Specifically, we model a node n_i's influence I_{n_i} as a bivariate gaussian distribution $N(L_{n_i}, \Sigma_{n_i})$, centered at n_i's location $L_{n_i} = (\alpha_{n_i}, \beta_{n_i})$ and with the covariance matrix Σ_{n_i} as its influence scope.

$$I_{n_i} = N(L_{n_i}, \Sigma_{n_i}) \tag{2}$$

We assume the influence scope of a node on the latitude and longitude dimensions is the same, so $\sum_{n_i} = \begin{pmatrix} \sigma_{n_i} & 0 \\ 0 & \sigma_{n_i} \end{pmatrix}$.

The influence probability of node n_i at a location L is measured in Equation 3:

$$P(L|I_{n_i}) = \frac{1}{\sigma_{n_i}^2} e^{\frac{(\alpha_{n_i}-\alpha_L)^2+(\beta_{n_i}-\beta_L)^2}{-2{\sigma_{n_i}}^2}} \tag{3}$$

2) Social User Influence Model: The probability that a user u_i influence a user u_j to build a following edge to him is measured in Equation 4:

$$P(f\langle u_j, u_i\rangle) = \frac{ST_{ji}}{\sigma_{u_i}^2} e^{ST_{ji}\frac{(\alpha_{u_i}-\alpha_{u_j})^2+(\beta_{u_i}-\beta_{u_j})^2}{-2{\sigma_{u_i}}^2}} \tag{4}$$

Equation 4 represents that a user with larger influence scope will attract more followers, a user will attract followers who are near to them easier and two users will have higher probability to be friends if they have more common friends.

3) Venue Influence Model: The probability that a user u_i check in at venue v_j is measured in Equation 5:

$$P(c\langle u_i, v_j\rangle) = \frac{1}{\sigma_{v_j}^2} e^{\frac{(\alpha_{v_j}-\alpha_{u_i})^2+(\beta_{v_j}-\beta_{u_i})^2}{-2{\sigma_{v_j}}^2}} \tag{5}$$

Equation 5 represents that a venue with larger influence scope will attract more users to check in and a venue will attract more users nearby to check in.

The probability that a user u_i rate venue v_j is measured in Equation 6:

$$P(r\langle u_i, v_j\rangle) = \frac{1}{\sigma_{v_j}^2} e^{\frac{(\alpha_{v_j}-\alpha_{u_i})^2+(\beta_{v_j}-\beta_{u_i})^2}{-2{\sigma_{v_j}}^2}} \tag{6}$$

Equation 6 represents that a venue with larger influence scope will attract more users to rate and a venue will attract more users nearby to rate.

4) IME Model on LBSNs: We make a conditional independence assumption that edges are conditionally independent given the head node and tail node. This assumption is widely applied in machine learning models like Naive Bayes[13]. For a LBSN, there are three types of edges: following edges, check-in edges and rating edges. We measure the weights of following edges, check-in edges and rating edges as W_f, W_c and W_r respectively. IME Model is shown in Equation 7 which measures joint probability of generating edges in LBSNs.

$$P(E|I_U, I_V) = \prod_{f\langle u_j,u_i\rangle\in F} p^{W_f}(f\langle u_j, u_i\rangle) \times \prod_{c\langle u_i,v_j\rangle\in C} p^{W_c}(c\langle u_i, v_j\rangle) \times \prod_{r\langle u_i,v_j\rangle\in R} p^{W_r}(r\langle u_i, v_j\rangle) \tag{7}$$

$$\sigma_{u_i}^2 = \frac{\sum_{u_j\in I_f(u_i)} ST_{ji}((\alpha_{u_j}-\alpha_{u_i})^2+(\beta_{u_j}-\beta_{u_i})^2)}{2|I_f(u_i)|} \tag{10}$$

$$\sigma_{v_j}^2 = \frac{\sum_{u_i\in(I_c(v_j)\cup I_r(v_j))} W_c(\alpha_{u_i}-\alpha_{v_j})^2 + W_r(\alpha_{u_i}-\beta_{v_j})^2}{2(W_c|I_c(v_j)| + W_r|I_r(v_j)|)} \tag{11}$$

$$\alpha_{u_i} = \frac{\sum_{u_j \in I_f(u_i)} W_f \frac{ST_{ji}\alpha_{u_j}}{\sigma^2_{u_i}} + \sum_{u_j \in O_f(u_i)} W_f \frac{T_{ij}\alpha_{u_j}}{\sigma^2_{u_j}} + \sum_{v_j \in O_c(u_i)} W_c \frac{\alpha_{v_j}}{\sigma^2_{v_j}} + \sum_{v_j \in O_r(u_i)} W_r \frac{\alpha_{v_j}}{\sigma^2_{v_j}}}{\sum_{u_j \in I_f(u_i)} W_f \frac{ST_{ji}}{\sigma^2_{u_i}} + \sum_{u_j \in O_f(u_i)} W_f \frac{T_{ij}}{\sigma^2_{u_j}} + \sum_{v_j \in O_c(u_i)} W_c \frac{1}{\sigma^2_{v_j}} + \sum_{v_j \in O_r(u_i)} W_r \frac{1}{\sigma^2_{v_j}}} \quad (8)$$

$$\beta_{u_i} = \frac{\sum_{u_j \in I_f(u_i)} W_f \frac{ST_{ji}\beta_{u_j}}{\sigma^2_{u_i}} + \sum_{u_j \in O_f(u_i)} W_f \frac{T_{ij}\beta_{u_j}}{\sigma^2_{u_j}} + \sum_{v_j \in O_c(u_i)} W_c \frac{\beta_{v_j}}{\sigma^2_{v_j}} + \sum_{v_j \in O_r(u_i)} W_r \frac{\beta_{v_j}}{\sigma^2_{v_j}}}{\sum_{u_j \in I_f(u_i)} W_f \frac{ST_{ji}}{\sigma^2_{u_i}} + \sum_{u_j \in O_f(u_i)} W_f \frac{T_{ij}}{\sigma^2_{u_j}} + \sum_{v_j \in O_c(u_i)} W_c \frac{1}{\sigma^2_{v_j}} + \sum_{v_j \in O_r(u_i)} W_r \frac{1}{\sigma^2_{v_j}}} \quad (9)$$

V. Home Location Identification method

In this section, we develop our Home Location Identification method based on IME model. We estimate unknown home location of users using the Maximum Likelihood Estimation(MLE) principle under IME model and the Home Location Identification problem is converted to optimize Equation 12. To be specific, we estimate users' home locations to maximize the logarithm of the likelihood which represents joint probability of generating edges(friendships, check-ins, ratings).

$$\max_{I_U, I_V} \log P(E|I_U, I_V) \quad (12)$$

We differentiate Equation 7 with regard to unknown variable and obtain the results shown in Equation 8, 9, 10, 11. In these equations, the unknown variables are dependent on each other. HLGPS uses a two-stage global iteration algorithm which is demonstrated in Algorithm 1 to solve the problem. In Stage 1, HLGPS initializes home location of users who have check-in data by clustering their check-in data using a sing-pass clustering algorithm. In Stage 2, HLGPS updates home location of users iteratively so that the likelihood is maximum.

A. Stage 1: Initializes home locations of users in U^{-H}

HLGPS initializes home location of users who don't have home locations from Step 3 to Step 9. For a user who has check-in data, HLGPS initializes his home location by clustering his check-in data C_{u_i} using a sing-pass clustering algorithm for locations called LocClustering inspired from the Single-pass Clustering Algorithm[5]. For a user who don't have check-in data, HLGPS initialize his home location as random. LocClustering clusters a location list to clusters in a single pass and returns the center of the largest cluster as result. The distance between a cluster and a location is defined as the distance between the location of the cluster center and the location. Specifically, LocClustering scans each location L_i in location list sequentially and find the nearest cluster C_{min} for current location L_i. If there are clusters and the distance d_{min} between the nearest cluster C_{min} and L_i is less than a threshold d_τ, it adds the location L_i to C_{min}. If there are no clusters or d_{min} is larger than d_τ, it creates a new cluster C_{new} with the location L_i. We denote the average number of check-ins of users as $\bar{c}$, then the complexity of Stage 1 is $O(|U^C| \times \bar{c})$ where $|U^C|$ is the number of users who have check-in data. Consequently, LocClustering is a linear algorithm.

B. Stage 2: Updating home locations of users iteratively

HLGPS updates home location of users who don't have check-in data iteratively from Step 12 to Step 30. The outer loop from Step 12 to Step 30 updates $\sigma^2_{u_i}$ and $\sigma^2_{v_j}$ based on Equation 10 and 11. The inner loop from Step 20 to Step 28 updates α_{u_i} and β_{u_i} based on Equation 8 and 9. HLGPS stops when the likelihood converges.

Algorithm 1 HLGPS

Input: $G, F, C, R, L_{u_i}(\forall u_i \in U^H), c_\tau$
Output: $L_{u_i}(\forall u_i \in U^{-H})$

1: **function** $HLGPS(G, F, C, R, L, c_\tau)$ ▷ HLGPS algorithm
2: // Stage 1: Init home location of users in U^{-H}
3: **for** each $u_i \in U^{-H}$ **do** ▷ users: no home location
4: **if** $u_i \in U^C$ **then** ▷ user: have check-in
5: $L_{u_i} = LocClustering(C_{u_i}, d_\tau)$
6: **else** ▷ user: no check-in
7: L_{u_i} = (random latitude, random longitude)
8: **end if**
9: **end for**
10: // Stage 2: Update home locations of users in U^{-H} iteratively
11: //Outer Loop
12: **while** true **do**
13: **for** each $u_i \in U$ **do**
14: Update $\sigma^2_{u_i}$
15: **end for**
16: **for** each $v_j \in V$ **do**
17: Update $\sigma^2_{v_j}$
18: **end for**
19: // Inner Loop
20: **while** true **do**
21: **for** each $u_i \in (U^{-H} \cap U^{-C})$ **do**
22: Calculate $\alpha^{new}_{u_i}$ and $\beta^{new}_{u_i}$
23: **end for**
24: If Inner Loop converges, then break
25: **end while**
26: **for** each $u_i \in (U^{-H} \cap U^{-C})$ **do**
27: $\alpha_{u_i} = \alpha^{new}_{u_i}$, $\beta_{u_i} = \beta^{new}_{u_i}$
28: **end for**
29: If Outer Loop converges, then break
30: **end while**
31: **end function**

VI. Experiments

A. Experiment Setup

The code[6] of HLGPS and dataset[7] are available online.

1) Dataset: In the dataset, there are 138,983 users who have check-in data, constituting only 16.7% of all users. For users who have check-in data, the average check-in number of each user is about 2.7. We define the ratio of people who have home location as rh and $rh = \frac{U^H}{U}$. In the experiments, we randomly split users into two parts: rh of users who have home location

[6]https://github.com/guyulongcs/HLGPS
[7]https://drive.google.com/open?id=0B7QrSjWN1vvYcm1NT3pCdXkxcDg

and $1-rh$ of users who don't have have home location. We set $rh = 80\%$ which is the same way as existing methods[1, 3, 9]. In this setting, there are 669,472 users have home location and 166,424 users who don't have home location. There are 27,781 users(16.7%) who have check-in data among the 166,424 users who don't have home location.

2) Methods:

- UDI is the method developed in [9], which predicts a user's location based on an influence model. UDI uses signals like friendships and venues in tweets.
- $Maxvote$ is the baseline method developed in [15], which predicts a user's location by taking the most popular location of a user. We can't directly using a max vote scheme because location information like latitude and longitude are continuous. So we firstly map check-in list to city list using method described previously.
- $ClusterHier$ is the baseline method developed in [10], which predicts a user's home location using a hierarchical clustering algorithm to cluster checkins at night(from 8:00 p.m. to 7:59 a.m.).
- Avg is the baseline method developed in [4, 17], which discretizes the world into 25 by 25 km cells and defines the home location as the average position of check-ins in the cell with the most check-ins.
- $HLGPS$ is our Home Location Identification method.
- $HLGPS_{-r}$ is our Home Location Identification method, but doesn't use rating data.
- $HLGPS_{tj}$ is our Home Location Identification method, but uses Jaccard Similarity to measure social trust.
- $HLGPS_{uc}$ is our Home Location Identification method, but also update users who have check-in data in the iteration stage.

3) Evaluation Metrics: We measure the performance of different methods using accuracy within 100 miles error distance(ACC) the same as previous work[9]. To be specific, for a user u_i, his true and estimated home location are L_{u_i} and $\widetilde{L}_{u_i}$ respectively. Let $Err(u_i)$ be the error distance between L_{u_i} and $\widetilde{L}_{u_i}$, then ACC is defined as Equation 13.

$$ACC = \frac{|u_i \in U^{-H} \wedge Err(u_i) \leq 100|}{|U^{-H}|} \tag{13}$$

B. Experiment Results

1) Home Location Identification on $U^{-H} \cap U^C$: Methods $Maxvote$, $ClusterHier$ and Avg have the shortcoming that they can only predict home locations of users who have check-in data. It means that they can only predict 16.7% of users in U^{-H} in the dataset. We firstly compare the performance of different methods on users who have check-in data($U^{-H} \cap U^C$). The performance of methods for Home Location Identification on $U^{-H} \cap U^C$ is shown in Figure 2. The result demonstrates that our method $HLGPS$ outperforms all existing methods for users who have check-in data. To be specific, $HLGPS$ can predict home locations of users who have check-in data at the accuracy 92.3% though the average check-in number of each user is only about 2.7, and achieves the best performance.

2) Home Location Identification on U^{-H}: In this experiment, we compare the performance of methods for all users who don't have home location(U^{-H}). The performance of methods for home location identification on all users who don't have home location is shown in Figure 4.

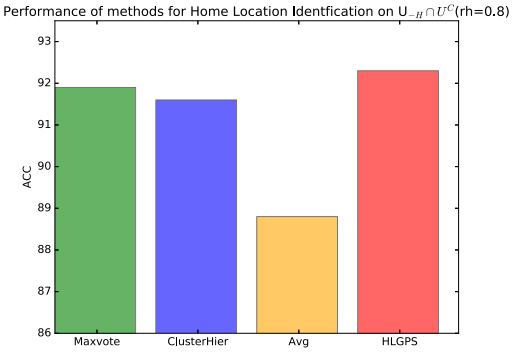

Fig. 2. Performance of methods on $U^{-H} \cap U^C$

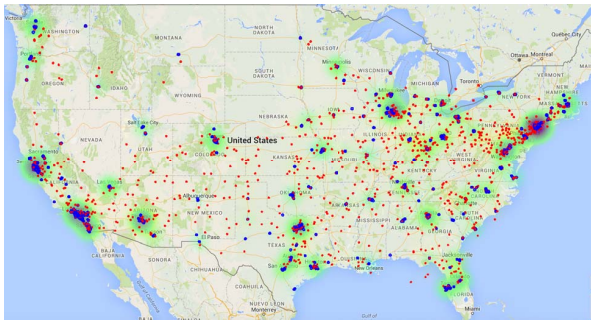

Fig. 3. Predicted and true home locations(red: predicted, blue: true)

a) $HLGPS$: We used grid search and found that $HLGPS$ has the highest accuracy of 67.7% when $W_f = 1, W_c = W_r = 0.1$. The predicted and true home location of users who don't have home location(U^{-H}) are shown in Figure 3.

b) $HLGPS$ vs. UDI: $HLGPS$ significantly improves UDI by 14.7% in terms of ACC.

c) $HLGPS$ vs. $HLGPS_{tj}$: By comparing $HLGPS$ and $HLGPS_{tj}$, we see that for the measurement of trust, sigmoid function improves ACC by 4.2% comparing to Jaccard Similarity.

d) $HLGPS$ vs. $HLGPS_{-r}$: By comparing $HLGPS$ and $HLGPS_{-r}$, we see that rating data can improve ACC by 3.5%. This demonstrates the effectiveness of using rating data for Home Location Identification problem.

e) $HLGPS$ vs. $HLGPS_{uc}$: By comparing $HLGPS$ and $HLGPS_{uc}$, we see that only update locations of users in U^{-C} in updating stage of $HLGPS$ can improve the ACC by 5.4%. In the initialization stage of HLGPS, we initialize home location of users in U^{-C} as random value and home location of users in U^C by clustering their check-in data. If we update home location of users in U^C using the randomly initialized locations of U^{-C} in the updating stage, the accuracy of estimated home location of users in U^C may be affected.

3) Influence of ratio of users who have home location: To investigate the influence of ratio of users who have home

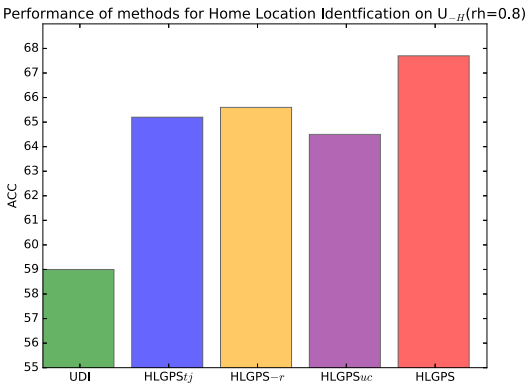

Fig. 4. Performance of methods on U^{-H} when $rh = 0.8$

location, we evaluate methods in another setting where $rh = 0.2$, which means that only 20% users have home location. In this setting which is more close to the real-world, $HLGPS$ has the accuracy of 55.9%, significantly outperforms UDI by 71.5%. We find that $HLGPS$ outperforms UDI even more when more users don't have home location.

C. Discussion

Content based approach [2, 3, 8, 11, 12] infers home location based on texts in social networks. This approach needs texts data in social network. What's more, venue information in texts can be noisy and ambiguous: Our method avoids problems like these using check-in data. Check-in based approach[4, 9, 10, 15, 17] infers home location of users using check-in data. Existing methods like $Maxvote$[15], $ClusterHier$[10] and Avg[4, 17] have the shortcoming that they can only predict home locations of users who have check-in data. Comparing to previous research, we firstly demonstrate the effectiveness of using social trust which measures closeness in social structure and rating data for the Home Location Identification problem.

VII. Conclusion and Future work

Home Location Identification of users in Location-based Social Networks is significant for location-based applications such as personalized search and recommendations. In this paper, we propose a Home Location Global Positioning System called HLGPS to solve the problem. Extensive experiments on a large scale dataset demonstrate that HLGPS outperforms state-of-the-art methods by 14.7%. In future, we will further study how to use time information for the Home Location Identification problem. What's more, we plan to do research on how to improve location-based services based on our Home Location Global Positioning System HLGPS and study how to protect privacy of users in social networks.

References

[1] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[2] Swarup Chandra, Latifur Khan, and Fahad Bin Muhaya. Estimating twitter user location using social interactions–a content based approach. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 838–843. IEEE, 2011.

[3] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010.

[4] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[5] William B Frakes and Ricardo Baeza-Yates. Information retrieval: data structures and algorithms. 1992.

[6] Justin J Levandoski, Mohamed Sarwat, Ahmed Eldawy, and Mohamed F Mokbel. Lars: A location-aware recommender system. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 450–461. IEEE, 2012.

[7] Michael Levandowsky and David Winter. Distance between sets. *Nature*, 234(5323):34–35, 1971.

[8] Guoliang Li, Jun Hu, Jianhua Feng, and Kian-lee Tan. Effective location identification from microblogs. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 880–891. IEEE, 2014.

[9] Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1023–1031. ACM, 2012.

[10] Hao Liu, Yaoxue Zhang, Yuezhi Zhou, Di Zhang, Xiaoming Fu, and KK Ramakrishnan. Mining checkins from location-sharing services for client-independent ip geolocation. In *INFOCOM, 2014 Proceedings IEEE*, pages 619–627. IEEE, 2014.

[11] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? inferring home locations of twitter users. *ICWSM*, 12:511–514, 2012.

[12] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):47, 2014.

[13] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[14] Ben Niu, Qinghua Li, Xiaoyan Zhu, and Hui Li. A fine-grained spatial cloaking scheme for privacy-aware users in location-based services. In *Computer Communication and Networks (ICCCN), 2014 23rd International Conference on*, pages 1–8. IEEE, 2014.

[15] Tatiana Pontes, Marisa Vasconcelos, Jussara Almeida, Ponnurangam Kumaraguru, and Virgilio Almeida. We know where you live: privacy characterization of foursquare behavior. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 898–905. ACM, 2012.

[16] Mohamed Sarwat, Justin J Levandoski, Ahmed Eldawy, and Mohamed F Mokbel. Lars*: a scalable and efficient location-aware recommender system. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2013.

[17] Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo. Socio-spatial properties of online location-based social networks. *ICWSM*, 11:329–336, 2011.