Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

From Interest to Function: Location Estimation in Social Media

Yan Chen[1,2], **Jichang Zhao**[1,3], **Xia Hu**[4], **Xiaoming Zhang**[1], **Zhoujun Li**[1] and **Tat-Seng Chua**[2]
[1]State Key Laboratory of Software Development Environment, Beihang University, P. R. China
[2]School of Computing, National University of Singapore, Singapore
[3]Department of Physics, Bar-Ilan University, Israel
[4]Department of Computer Science and Engineering, Arizona State University, United States

Abstract

Recent years have witnessed the tremendous development of social media, which attracts a vast number of Internet users. The high-dimension content generated by these users provides an unique opportunity to understand their behavior deeply. As one of the most fundamental topics, location estimation attracts more and more research efforts. Different from the previous literature, we find that user's location is strongly related to user interest. Based on this, we first build a detection model to mine user interest from short text. We then establish the mapping between location function and user interest before presenting an efficient framework to predict the user's location with convincing fidelity. Thorough evaluations and comparisons on an authentic data set show that our proposed model significantly outperforms the state-of-the-arts approaches. Moreover, the high efficiency of our model also guarantees its applicability in real-world scenarios.

Introduction

A rapid growth of online social media, especially Twitter (twitter.com) and Weibo (www.weibo.com), has provided Internet users a powerful and convenient means of updating status, sharing information and performing virtual social activities. For example, Twitter has over 300 million registered users and they publish over 140 million microblog posts, known as tweets, every day. Weibo has also accumulated more than 300 millions users in less than three years and more than 1,000 Chinese tweets are being posted per second (Zhao et al. 2012).

The high-dimension content generated by millions of users presents both opportunities and challenges to the contemporary research. Regarding the users as social sensors, we can collect tremendously large data set to facilitate the understanding of user behaviors (Song et al. 2010; Guimera et al. 2012). Part of previous efforts benefits from the knowledge of user location. For instance, the local news summarization from nearby Twitter users (Yardi and Boyd 2010), the location based recommendations (Cheng et al. 2012) for news (Phelan, McCarthy, and Smyth 2009) or activities (Zheng et al. 2010), extracting local news event (Agarwal et al. 2012), Twitter-based earthquake detection (Sakaki, Okazaki, and Matsuo 2010) and disease outbreak finding (Eubank et al. 2004). While at the same time, the content in the tweets are short, sparse or even noisy, particular for the location. It is found that only 26% of Twitter users list their location feature as granular as a city name (e.g.: California) in their profile, while the rest just fill them over general region or leave them blank or with nonsensical information (e.g.: Wonderland) (Cheng, Caverlee, and Lee 2010). Since August 2009, Twitter began to support the per-tweet geotagging (geo-tagged) function and Weibo also provides the similar feature, especially for the mobile platform. It seems that the social media platforms realize the need for a finedtuned user tracking by associating each tweet with a latitude and longitude. While in fact, only less than 1% of tweets are geo-tagged (Mahmud, Nichols, and Drews 2012), which limits the impact of those location-based sensing system.

To tackle the problem of location sparsity, many efforts have been devoted in the previous work. By identifying specific location-based words, Cheng et al. (2010) only used the content of a user's tweets to estimate the city-level location of the user. Sadilek et al. (2012) employed the word occurrence and other features to discover the potential users that have the same behavior pattern with the target user, and captured both directions of the relationship between location and social ties (Crandall et al. 2010). However, the potentially strong but inconspicuous relation between user interest and location function is overlooked. In fact, a tweet often reflects the posting user's interest or behavior, and given a location, it can generally be assigned with a functional semantic. Hence, a user with a certain interest is more plausible to be at the location with semantically similar function frequently. For example, given a user "Jean" who likes to go shopping, and if she tweets in the shopping mall once, then it is highly likely that she will post a tweet with similar contents or interest topic in future in the shopping mall, although she might not be in a same shopping place. Based on these observations, in this paper, for the task of estimating content-based location, instead of simply identifying location specific words or using co-occurrence of words, we

deeply explore users' location preference by users' interest, which leverages tweets content similarity, tweets content topic, users interaction behavior (comment, retweet and mention) in a higher and more abstract level. Combining with users' historical tweets with geo-tag, we construct the mapping between users' interest and the real physical locations or points of interest (POIs), and we use this to predict the user's location. We perform experiments on a large Chinese tweets data set, and demonstrate that our proposed approach yields better results as compared to the state-of-the-arts methods. Moreover, our model is simple and efficient, which can be applicable in real-world scenarios.

The main contributions of this research can be summarized as follows:

- To the best of our knowledge, this is the first attempt towards predicting content-based location through the mining of users' interests.
- Our work greatly enriches the semantics of locations, as it incorporates POI, which can reflect the function of location.
- Our approach constructs a bridge between users' interest and location, which can be further applied to many location based applications such as location-based recommendation ect.

Related Work

Generally in previous literature, the proposed methods about location estimation can be categorized into two classes. The common intuition of some representative works are that a user's tweets may encode some location-specific content - either specific place names or certain words or phrases more likely to be associated with certain locations than others. Cheng et al. (2010) built multi local term classifiers for identifying words in tweets with a strong geo-scope and estimate users' location within a probabilistic framework. Mahmud et al. (2012) constructed a time zone location classifier based on users' tweeting behavior and used an ensemble of statistical and heuristic classifiers to predict home location of Twitter users. Roller et al. (2012) performed text-similarity distance based method to estimate a test document location with weighted voting from top k most similar documents. As users related in social networks usually share common attributes, other representative works have been studying how a user's private information such as location could be inferred through an analysis of the users' social ties (Heatherly, Kantarcioglu, and Thuraisingham 2009; Lindamood et al. 2009; Gao, Tang, and Liu 2012). As one's total friends' number tends to decrease as the distance increases (Mok and Wellman 2007), Backstrom et al. (2010) predicted the home address of Facebook users based on provided addresses of their friends. However, the approach is only suitable for estimating users' home address, as there is not necessarily a connection between a users' dynamic location with their real friends most of the time (Cho, Myers, and Leskovec 2011). Sadilek et al. (2012) presented a system which implemented a probabilistic model of human mobility. But during the link prediction phase, for each user, Sadilek's method will compare with all the others users in the data set to find the top k users that have the similar mobile patterns. In order to overcome the problems of Sadilek et al. (2012) method and others, our work fully looks into users' interest and concentrates on location preference.

The problem of interest detection has been studied in a large number of domains, especially on the use of Bayesian probabilistic model for community discovery. The efforts mainly focus on discovering similar interest groups of people that are keen on talking about the same topic (Zhang et al. 2007; Henderson et al. 2010). Since the content of tweets could reflect users interest, recent research efforts have started to investigate methods that combine both tweets content and link information available in social networks (Zhou et al. 2006; Pathak et al. 2008; Sachan et al. 2011). However, different from the works that discover users interest group, we pay more attention on finding each user's interest distribution for preparation of our location prediction phase.

Method

Overview

In this section, we introduce the whole processing of our approach. It mainly comprises three phases: interest detection, mapping from location function to interest, and location estimation.

Interest Detection: It is worth noting that users' tweets can reflect their personal interest. Based on this observation, in this phase, we sort the tweets of each user in a chronology order, and employ a topic model to discover the hidden interest distribution.

Mapping from Location Function to Interest: Different locations have different functions for users. We could construct the hidden relationship between users' interest and the functions of real physical places.

Location Estimation: The users' activity scope is usually extremely limited, not far from their homes (Cho, Myers, and Leskovec 2011). Given the historical locations and user interest, we could establish a simple Bayesian model to predict the current location from the history records.

Interest Detection

In this paper, with the aim of handling the short-text tweet, we try to present a method of interest detection based on LDA Model (Blei, Ng, and Jordan 2003). In the standard LDA, a document contains a mixture of topics, represented by a topic distribution, and each word has a hidden topic label. While this is a reasonable assumption for long documents, for short microblog posts, a single post is most likely to be about a single topic (Diao et al. 2012; Chen et al. 2012). Therefore, we associate a single hidden variable for each post to indicate its topic. It is worthy to be noted that in the Twitter-like social media, there are three kinds of interactions between different users, including comment, retweet and mention. The frequencies of these behaviors could be utilized to indicate the strength of social ties between users or reflect the similarity of users' interest. Therefore we also import these interactions into our detection method.

Our interest detection model has three levels: for each user, the topic distribution θ and interest distribution i are generated; for each message, a topic z and an interaction type x are generated specific to the chosen topic z; finally, each word is generated specific to the chosen topic z. Figure 1 depicts the resulting probabilistic graph model, and the corresponding notations are summarized in Table 1. Overall, the generation process can be described as follows:

1) Sample $\phi_{u,z} \sim$ Dirichlet(β), $\varphi_{u,i} \sim$ Dirichlet(λ), $\eta_u \sim$ Dirichlet(γ);
2) For each user $u = 1, \ldots, U$,
   (a) Sample $\theta_u \sim$ Dirichlet(α_u);
   (b) Sample $i_u \sim$ Dirichlet(η_u);
3) For each post $t = 1, \ldots, T$,
   (a) sample a topic $z \sim$ Multinomial(θ_u);
   (b) sample an interaction type $x \sim$ Multinomial(i_u);
4) For each word $n = 1, \ldots, N$,
   sample $w_{m,n} \sim$ Multinomial ($\phi_{u,z}$).

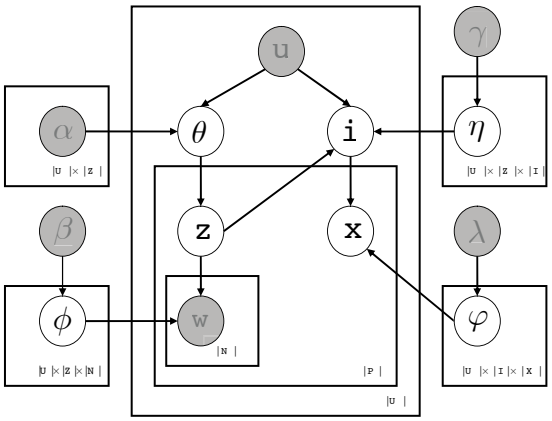

Figure 1: The probability graph model representation for interest detection model.

According to the probability graph model, the joint distribution of all variables can be written as:

$$\begin{aligned} J &= P(w, z, \theta, x, i \mid \alpha, \beta, \lambda, \gamma, u) \\ &= \sum_{\phi} \sum_{\eta} \sum_{\varphi} P(w, z, \theta, x, i, u, \phi, \varphi, \eta \mid \alpha, \beta, \lambda, \gamma) \\ &= \sum_{\phi} \sum_{\eta} \sum_{\varphi} P(w \mid z, u; \phi) P(z \mid \theta, u) P(\theta \mid \alpha, u) \\ &\quad P(x \mid i, u; \varphi) P(i \mid z, u; \eta) P(\eta \mid \gamma) P(\varphi \mid \lambda) P(\phi \mid \beta) \end{aligned} \tag{1}$$

Due to parameter couplings in the models, calculating exact posterior probabilities over all the hyper-parameters is intractable. Hence we use collapsed Gibbs sampling to obtain samples of the hidden variable assignment and estimate the model parameters from these samples. Gibbs sampling is carried out by starting with a random assignment to all the latent variables, using the update equations to compute fresh latent assignments over a large burn-in period. When this distribution stabilizes, sufficient number of samples are taken at regular intervals to avoid correlation. The Gibbs update equations are:

$$\phi = P(w \mid z, \neg w, \beta, u) = \frac{n^{-p}_{w,z,u} + \beta}{\sum_{v \in W} n^{-p}_{v,z,u} + W\beta} \tag{2}$$

$$\varphi = P(x \mid i, \neg x, \lambda, u) = \frac{n^{-p}_{x,i,u} + \lambda}{\sum_{r \in X} n^{-p}_{r,i,u} + X\lambda} \tag{3}$$

$$\eta = P(i \mid z, u, \neg i, \gamma) = \frac{n^{-p}_{i,z,u} + \gamma}{\sum_{s \in U} n^{-p}_{s,z,m} + W\gamma} \tag{4}$$

For each user, we get each interest distribution $p_{ui}(I)$, and each tweet is assigned an interest label of the user.

α,β,γ,λ	hyperparameters and priors of Dirichlet distributions.
θ	the $\|U\| \times \|Z\|$ matrix indicating user-topic distribution.
ϕ	the $\|U\| \times \|Z\| \times \|N\|$ matrix indicating user-topic-word distribution.
η	the $\|U\| \times \|Z\| \times \|I\|$ matrix indicating user-topic-interest distribution.
φ	the $\|U\| \times \|I\| \times \|X\|$ matrix indicating user-interest-interaction distribution.
u,i,z, x,w,t	the instance of a variable: u for user, i for interest, z for topic, x for interaction type, w for word, t for tweet.
U,I,Z, X,N T	users collection, interest collection, topic collection,interaction types collection, word collection, tweets collection in the dataset.
$n^{-p}_{j,m,u}$	the number of times j is generated from m for user u in the model, excluding post p.

Table 1: Important notations used in this paper and their descriptions.

Mapping from Location Function to User Interest

With the emergency of mobile social networks, location-based services have been available in many social medias. Diverse convenient interfaces have been provided to facilitate the users to check in, through which they could embed their real-time locations or nearby POIs into the tweets. Moreover, the POIs are divided into different classes for their different functional categories, including sports, entertainment, etc. Intuitively, we could employ POI as a bridge to construct a mapping between users' interest and location functions. We define the mapping as $P(C|I)$, where C stands for location function, I represents interest, its value stands for the probability of a user with interest I would like to go to the place with function C. We assume that the user interest is strongly related to the location with semantically similar function.

Location Estimation

Based on the previous stages, we could detect the interest distribution for a given user or tweet and obtain the mapping from location function to interest from the training data. Meanwhile, for many users, their everyday activity is limited

to a small radius, i.e., not far from there homes (Cho, Myers, and Leskovec 2011). And for most of them, several regular locations are frequently visited, like home, work place, shopping market, etc. Hence, we assume that the activity scope of most of users is not large, and the users with a large activity scope are quite few. We then can establish a simple but efficient Bayesian framework to predict the current location of a user based on the newly posted tweet.

For a user u, a joint probability of u and location l can be represented as equation (5):

$$P(l, u) = p(u)p(l|u). \tag{5}$$

We assume that $P(u)$ is a constant. Our object is to maximize the likelihood location y_u for a given posting by user u:

$$y_u = \arg\max_l P(l|u). \tag{6}$$

By employing the interest distribution of user u, y_u can be written as:

$$y_u = \arg\max_l P(l|u) = \arg\max_l P(l|I)P(I|u). \tag{7}$$

Furthermore, we import POI category as a bridge, and construct the interest and POI category relationship. We can obtain y_u as:

$$\begin{aligned} y_u &= \arg\max_l P(l|u) = \arg\max_l P(l|I)P(I|u) \\ &= \arg\max_l P(l|C, u)P(C|I)P(I|u). \end{aligned} \tag{8}$$

In summary, for a given user, we first detect the interest distribution and label all the tweets with interest. We can then obtain the location prediction based on the historical records and the function-interest mapping.

Experimental Results

This section presents the basic statistics of our data set as well as the construction of ground truth. Based on the data set, we first validate our assumptions in the above subsections. We introduce three performance metrics to evaluate our proposed method. Based on these metrics, we find that our approach outperforms the state-of-art models on geo-location prediction.

Dataset and Ground Truth

Our experiments are based on the Chinese tweets obtained from Weibo, the most popular micro-blogging service in China where people post messages with at most 140 characters. Weibo has published its APIs since 2010 and through these APIs, we collected 25,869,257 public Chinese tweets from June 2012 to July 2012. However, there are only 1,567,525 (accounts for 6%) tweets with geo-tagged in the form of latitude and longitude coordinates. This verifies that location information of tweets is extremely sparse. Since these tweets generally are posted by smartphone, we assume that these locations are correct and can be used as ground truth. To tackle the sparseness of dataset, we follow a similar strategy suggested by Cheng et al. (2010) and Sadilek et al. (2012). We further locate those tweets into Beijing corresponding city-level region, and arrive at 123,716 tweets and 90,732 users. We crawl those users' tweets history. Finally we arrive at 1,113,135 tweets posted by 312,857 users with geo location.

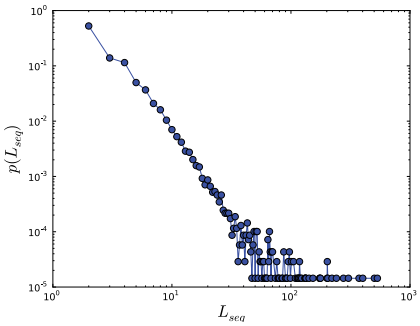

Figure 2: Distribution of the length of users' location sequences

We sequence each user's tweets in a time line order, and treat each user's latest tweet as testing data, while using the earlier ones as training data. We define the length of a user's tweets sequence as L_{seq} and plot its distribution in Figure 2. It can be found that the data set we used is very sparse, because most of the users only have less than ten tweets with geo-tags. Moreover, we collect 435,801 representative POIs in Beijing, and label all tweets' geo with the nearest POI [1].

Validation of the Assumptions

To begin with, we obtain each users' interest distribution based on our proposed interest detection model, where interest numbers are all set to 10. The choice of hyper-parameters α, β, λ and γ can have important implications for the results produced by the interest detection phase. In our dataset, we tend to employ a small value of interest number. Increasing β and λ can be expected to decrease the number of topics and interest. In other words, the tweets collection can be sensibly factorized into a set of topics and interest at several scales, and the particular scale of the topics and interest assessed by the model will be set by β and λ. Thus, we set the smaller values for β and λ, where $\beta = 0.05$, and $\lambda = 0.02$. We set $\alpha|Z| = constant$ and $\gamma|I| = constant$, topic number $|Z| = 20$, interest number $|I| = 10$, and $constant = 10$, thus we get $\alpha = 0.5$ and $\gamma = 1$.

Based on this, our assumptions presented in the previous section can be validated through the data set we employ. As shown in Figure 3, it can be learnt that semantically similar location function and user interest are strongly correlated. It can be clearly seen from the figure that for each function category (column), it obviously has an corresponding interest. For example, for category "Finance and assurance" (C_4), it has a high correlation with i_2. The similar case like category "Shopping" (C_9) to i_8, "Science and education" (C_{10}) to i_6, etc. For each learnt interest (row), it has some comparative strong correlated categories too. For

[1] http://open.weibo.com/wiki/2/place/pois/category

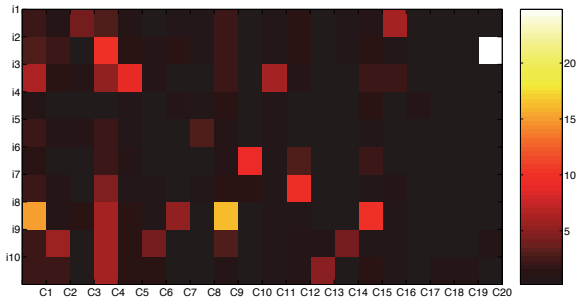

Figure 3: Function-interest mapping

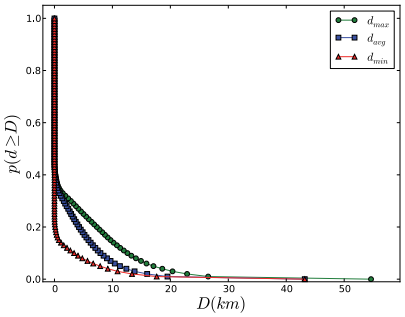

Figure 4: Activity radius distribution in the real world data Set.

example, for i_3, it has a comparative correlation with "Eating" (C_1), "Living" (C_5) and "Health" (C_{11}), meanwhile these three categories do have strong relationship in reality. Another case is for i_8, whose corresponding categories are "Eating" (C_1), "Shopping" (C_9), "Leisure" (C_{15}) and "Famous Place" (C_7). Moreover, we look into these interest and find the top ranked keywords of most of interest are meaningful, such as "bag", "car", "sale" "good", "sunshine" for i_8, "service" for i_2, etc.

Regarding the activity scope, we find that for the Weibo user, the radius of mobility is extremely limited. For each user, the location sequence is denoted as $\{l_i\}(i = 1, 2, ..., n)$, and we calculate the distance between l_n and other history locations in $\{l_i\}(i = 1, 2, ..., n-1)$ and obtain the distributions for maximum distance (d_{max}), averaged distance (d_{avg}) and minimum distance (d_{min}). As can be seen in Figure 4, most users' current location appears in the history locations, i.e., $d_{min} = 0$. It supports our conjecture that the current location can be predicted by selecting the most plausible location from the history.

Measurements

In this work, we utilize three widely used metrics to evaluate our geo prediction task: error distance ($ErrDist$), average error distance ($AverErrDist$) and accuracy ($Accuracy$). For a user u, let $ErrDist$ be the error distance between a user's actual location and an estimated location, as defined in equation (9),

$$ErrDist(u) = d(l_{act}(u), l_{est}(u)). \quad (9)$$

Table 2: $Accuracy$ and $AverErrDist$ (AED) performance with varying ground truth generation

K	0.5km	1km	1.5km	2km	3km
Accuracy	0.67	0.72	0.58	0.45	0.28
AED	0.734	0.8686	2.397	4.826	10.302
#Tweets	135,347	244,415	281,843	295,871	313,059
#Users	48,425	65,934	76,328	78,974	82,447

For a set of users U, $AverErrorDist$ means to what extent the approach can geo-locate users close to their actual location on average, and $Accuracy$ considers the percentage of users with their error distance less than or equal to a specific threshold ($Thresh$), which are defined in equation (10) and (11) respectively.

$$AverErrDist(U) = \frac{\sum_{u \in U} ErrDist(u)}{|U|}, \quad (10)$$

$$Accuracy(U) = \frac{\sum_{u \in U} ErrDist(u) \leq Thresh}{|U|}, \quad (11)$$

where $l_{act}(u)$ is the actual location of the user u and $l_{est}(u)$ is the estimated location of u. In the following evaluations, we set $Thresh = 0km$, which is a rigid but convincing value for our location estimation tasks.

Performance Analysis

For each tweet with geo-tag, we need to find a POI to represent its location function. In the following experiments, we find a closest POI within the radius K. If this failed, the tweet without proper POI would be omitted from the data set. We first conduct experiments to evaluate the effectiveness of our proposed method with progressively increasing K.

Table 2 displays the accuracy, average error distance (AED), the tweets and users numbers with varying value of K. It is observed that our proposed method achieves remarkable accuracy and AED for K less than $1.5km$. As K is less than this value, the users interest generally have a strong relationship with POI location. As K becomes large, especially when K equals to $3km$, our approach can not produce good $Accuacy$ and AED performance. This poor performance is mainly due to the unreliable pseudo ground truth which brings some noise onto our approach; in another word, the user interest mined in this case may not have a strong correlation with POI. Therefore this observation verifies that if the users' interest have a strong correlation with the POI, our approach can achieve a promising performance. The proportion of effective data can also influence the $Accuracy$ and AED performance. Most of the data for K equals to $0.5km$ and $1km$ are both very effective. In fact, in order to avoid the influence of tweets number for each user when we set different K values, we make sure that the average number of tweets of each user does not vary dramatically, although the number of tweets with geo location and users in our data set are increasing with the growth of the distance threshold K.

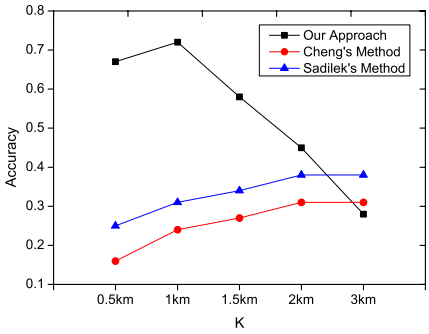

Figure 5: Accuracy

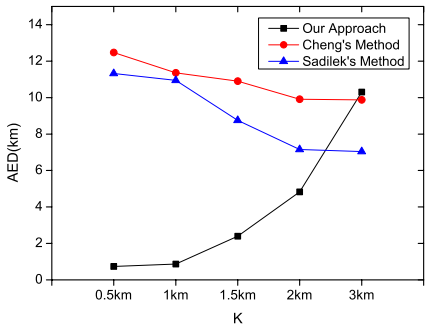

Figure 6: AED

Comparison with other methods

To fully demonstrate the effectiveness of our proposed approach, we compare it against the following state-of-the-art methods, which are most similar to ours.

- **Cheng's Method** (Cheng, Caverlee, and Lee 2010) assigns a location to a Twitter user based on a set of local words identified by the tweets content. In our experiment, we first use their proposed method to filter out the local words, then the strong smoothing method (Lattice-based Neighborhood Smoothing) is employed as the baseline.
- **Sadilek' Method** (Sadilek, Kautz, and Bigham 2012) utilize tweets content and location to discover the potential relationship of users. Based on friendship graph, they predict user location. In our experiment, we first employ two features (text similarity, co-location) to estimate users friendship, the goal of which is to discover users with similar activity behavior and construct the relation between them. We then predict users location based on the entire graph through the Hidden Markov Model.

Figures 5 and 6 illustrate the $Accuracy$ and $AverErrDist$ performance with various values of K, respectively. It is observed that our approach is generally significantly better than the two representative methods. It can be seen from Figure 5 that our approach produces a higher accuracy than the two methods at the smaller value of K. This verifies that users' interest and location preference mining plays a positive role in location estimation. While as K increases, its accuracy quickly decreases, and at some points, it drops to almost the same performance as the two methods, or even worse. This is due to the fact that our method mainly depends on users' history location information at this point, as the POI almost has no relationship with users location, or even worse, the error POI category may produce a negative impact for location estimation.

It is worth noting from Figure 6 that the Average Error Distance for Sadilek's method decreases quickly when the values of K are greater than $0.5km$. This is because their method are keen on discovering users with similar patterns; however, when the user number increases, it is more likely to discover users with the similar behavior patterns.

Low time complexity of the prediction algorithm is also fundamental, especially for analyzing large data set. For two baselines and our proposed method, most of the time is spent during off-line preprocessing phase before the prediction. For example, the local words discovery in Cheng's method, link prediction for Sadilek's method, and interest detection in our approach. Although during the on-line phase, these three methods all perform efficiently, if there is a new user, Sadilek's method needs to compare the new comer with all the existing ones in the dataset, which would consume a lot of time and makes this method slow. While for our model, it can generate the users interest through trained interest detection model in almost real time.

Conclusion and Future Work

Many research efforts have been devoted to user location estimation in recent years. In this paper, we proposed an efficient three phases location estimation approach to estimate user's location based on tweets purely, which addresses two concerns. On one hand, we deeply looked into tweets content and mined the users' interest and location preference. On the other hand, our approach utilized POI, which greatly enriches the function semantic of predicted geo-locations. The experimental results demonstrated its effectiveness and applicability as compared to existing methods.

Although the model presented in this paper is independent of the language, it might be affected by the culture or user behaviors in different regions. Hence in the future work, we would like to try on more data sets from different countries to perform more thorough evaluations. Moreover, we would like to further investigate users' behavior and real word location preference to recommend potential locations to users.

Acknowledgments

The authors thank the anonymous reviewers for their constructive suggestions. This work was supported by the National Natural Science Foundation of China (60973105, 90718017, 61170189, and 61202239), the Research Fund for the Doctoral Program of Higher Education (20111102130003), the Fund of the State Key Laboratory of Software Development Environment (KLSDE-2011ZX-03) and the Singapore National Research Foundation and Interactive Digital Media R&D Program Office, MDA under research grant (WBS:R-252-300-001-490). JZ thanks the China Scholarship Council (CSC) for support.

References

Agarwal, P.; Vaithiyanathan, R.; Sharma, S.; and Shroff, G. 2012. Catching the long-tail: Extracting local news events

from twitter. In *Proceedings of Association for the Advancement of Artificial Intelligence*.

Backstrom, L.; Sum, E.; and Marlow, C. 2010. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the International Conference on World Wide Web*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Chen, Y.; Li, Z.; Nie, L.; Hu, X.; Wang, X.; Chua, T. S.; and Zhang, X. 2012. A semi-supervised bayesian network model for microblog topic classification. In *Proceedings of the International Conference on Computational Linguistics*.

Cheng, C.; Yang, H.; King, I.; and Lyu, M. R. 2012. Catching the long-tail: Extracting local news events from twitter. In *Proceedings of Association for the Advancement of Artificial Intelligence*.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of ACM Conference on Information and Knowledge Management*.

Cho, E.; Myers, S. A.; and Leskovec, J. 2011. Friendship and mobility: User movement in location-based social networks. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining*.

Crandall, D. J.; Backstrom, L.; Cosley, D.; Suri, S.; Huttenlocher, D.; and Kleinberg, J. 2010. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences* 107:22436–22441.

Diao, Q.; Jiang, J.; Zhu, F.; and Lim, E. P. 2012. Finding bursty topics from microblogs. In *Proceedings of Association for Computational Linguistics*.

Eubank, S.; Guclu, H.; Anil Kumar, V. S.; Marathe, M. V.; Aravind, S.; Toroczkai, Z.; and Wang, N. 2004. Modelling disease outbreaks in realistic urban social networks. *Nature* 429:180–184.

Gao, H.; Tang, J.; and Liu, H. 2012. Exploring social-historical ties on location-based social networks. In *Proceedings of Association for the Advancement of Artificial Intelligence*.

Guimera, R.; Llorente, A.; Moro, E.; and Sales-Pardo, M. 2012. Predicting human preferences using the block structure of complex social networks. *PLoS ONE* 7:e44620.

Heatherly, R.; Kantarcioglu, M.; and Thuraisingham, B. 2009. Social network classification incorporating link type. In *Proceedings of IEEE Intelligence and Security Informatics*.

Henderson, K.; Eliassi-Rad, T.; Papadimitriou, S.; and Faloutsos, C. 2010. Hcdf: A hybrid community discovery framework. In *Proceedings of the SIAM International Conference on Data Mining*.

Lindamood, J.; Heatherly, R.; Kantarcioglu, M.; and Thuraisingham, B. 2009. Inferring private information using social network data. In *Proceedings of the International Conference on World Wide Web*.

Mahmud, J.; Nichols, J.; and Drews, C. 2012. Where is this tweets from? Inferring home location of twitter users. In *Proceedings of Association for the Advancement of Artificial Intelligence*.

Mok, D., and Wellman, B. 2007. Did distance matter before the internet? interpersonal contact and support in the 1970s. *Social networks* 29:430–461.

Pathak, N.; DeLong, C.; Banerjee, A.; and Erickson, K. 2008. Social topics models for community extraction. In *Proceedings of the 2nd SNA-KDD Workshop*.

Phelan, O.; McCarthy, K.; and Smyth, B. 2009. Using twitter to recommend real-time topical news. In *Proceedings of ACM conference on Recommender systems*.

Roller, S.; Speriosu, M.; Rallapalli, S.; Wing, B.; and Baldridge, J. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of International Conference on Empirical Methods in Natural Language Processing*.

Sachan, M.; Contractor, D.; Faruquie, T. A.; and Subramaniam, L. V. 2011. Probabilistic model for discovering topic based communities in social networks. In *Proceedings of ACM Conference on Information and Knowledge Management*.

Sadilek, A.; Kautz, H.; and Bigham, J. P. 2012. Finding your friends and following them to where you are. In *Proceedings of ACM Conference on Web Search and Data Mining*.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: real time event detection by social sensors. In *Proceedings of the International Conference on World Wide Web*.

Song, C.; Qu, Z.; Blumm, N.; and Barabasi, A. L. 2010. Limits of predictabilility in human mobility. *Science* 327:1018–1021.

Yardi, S., and Boyd, D. 2010. Tweeting from the town square: Measuring geographic local networks. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Zhang, H.; Giles, C. L.; Foley, H. C.; and Yen, J. 2007. Probabilistic community discovery using hierarchical latent gaussian mixture model. In *Proceedings of Association for the Advancement of Artificial Intelligence*.

Zhao, J.; Dong, L.; Wu, J.; and Xu, K. 2012. Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining*.

Zheng, V. W.; Zheng, Y.; Xie, X.; and Yang, Q. 2010. Collaborative location and activity recommendations with GPS history data. In *Proceedings of the International Conference on World Wide Web*.

Zhou, D.; Manavoglu, E.; Li, J.; Giles, C. L.; and Zha, H. 2006. Probabilistic models for discovering e-communities. In *Proceedings of the International Conference on World Wide Web*.