Graph-Based Residence Location Inference for Social Media Users

Dan Xu, Peng Cui, Wenwu Zhu, and Shiqiang Yang
Tsinghua University, China

By jointly considering social, visual, and textual information for geographically nearby social media users, the proposed framework can infer missing residence location information.

As the amount of location information increases, location becomes one of the most vital attributes of a social media user, one that can bridge the gap between the user's online and offline activities. If all social media information were geolocated, then social media could play an important role in addressing serious and significant real-world problems, such as election prediction,[1] epidemic forecasting,[2] and emergency detection.[3]

However, the disclosure of location raises serious privacy and security concerns. Users providing location-specific status updates and/or sharing locations may make it possible for attackers to identify their trajectories, possibly causing serious security problems. Because of such privacy and security concerns, most users are unwilling to publish or exactly describe their locations in social media. Therefore, it is only possible to identify users' locations from public information, such as user profiles and social relationships.

For the purposes of this discussion, we classify location into residence and current. *Residence location* is defined as the city-level place where most of a user's activities occur. It captures the user's long-term geographic range rather than a real-time spatial point. Although *current location* has a finer granularity, city-level residence is as sufficient as current location for many important applications. Also, because most social media content is generated at a user's residence location, it is effective and efficient to use it to approximate the locations of most user-generated content. Many location-based applications tend to leverage residence location as a symbolic feature. Therefore, in this work, we focus on users' residence locations in social media.

To predict a user's residence location, most prediction methods leverage user-generated content (content-based approach), social relationships (social-based approach), or a combination of both methods (combined approach). Content-based methods predict locations by identifying location words in user-generated content. These methods do not perform well with social media because of the weak relationship between the user's true geographic position and the location mentioned in the content. Moreover, high computation costs result in poor scalability of these methods. Social-based approaches assume that friends in social media are located near each other, so they leverage the user's social relationships to predict locations for unknown users. These two lines of research are orthogonal and complement each other. Thus, jointly considering the content and social information has the potential to more optimally infer users' residence locations. (See the "Related Work in Location Prediction" sidebar for earlier work on this topic.)

This line of work still faces several challenges, however. First, given that only a few users' residence locations are known (data sparsity), how do we make full use of that data to infer the unknown locations for the majority of users? Second, users' following behaviors and their generated content are often casual and uncertain, resulting in noisy data. Therefore, we cannot directly apply our existing prior to solve the sparsity problem. Finally, both social and content data contain heterogeneous information, and they play different roles in predicting residence locations. How do we balance them?

Considering all these challenges, we propose a novel framework for residence location

Published by the IEEE Computer Society

Related Work in Location Prediction

In recent years, there has been extensive research in the location prediction field, particularly work focused on social media. This active research area can be differentiated into three primary types: location categories, inference techniques, and spatial sources.

Various location prediction methods can have different goals. Some approaches predict users' residence locations, some infer the location a user is talking about, and others focus on where a tweet was posted. The use of location information differs for each goal. For example, the location mentioned in a tweet may be leveraged for disaster detection or emergency response, the location where a tweet was posted could be used by a location-based service, and a user's residence location could be utilized by a personal service or market survey.

Most inference technique research can be broadly classified into two approaches that use either natural language processing (NLP) or gazetteers. NLP-based approaches leverage language models, and gazetteer-based approaches use a toponymy dictionary to determine location. The gazetteer-based approach does not require training data, but it often performs poorly if the content is not in the gazetteer. NLP-based approaches may leverage gazetteers as a baseline or to train models.

Different spatial information sources are used for location inference, and location prediction research can be broadly divided into three categories: content-based, social-based, and combined approaches.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee focused on the residence location inference in Twitter by leveraging the local terms posted in a specific geographic region.[1] Swarup Chandra, Latifur Khan, and Fahad Bin Muhaya developed a language model based on users' conversations.[2] In their model, all terms in the same conversation belong to the conversation initiator. Hau-wen Chang and his colleagues inferred user locations without training data by proposing the location distributions of terms based on a Gaussian mixture model.[3] Their experiments confirmed that the method achieves better accuracy. Unlike these methods utilizing user-generated content, our method employs social relationships and textual and visual content, and it works independent of language.

Lars Backstrom, Eric Sun, and Cameron Marlow introduced a location estimation method for Facebook using probabilistic inference based on a user's friends.[4] They first assigned the probability of friendship given users' geographic distances and then evaluated a user's location by employing maximum likelihood estimation. Adam Sadilek, Henry Kautz, and Jeffrey P Bigham predicted users' trajectories based on social relationships.[5] These social-based approaches assume users at the same distance have the same probability of friendship. In fact, this is usually invalid. Therefore, these models cannot differentiate users with different influence.

Eunjoon Cho, Seth Myers, and Jure Leskovec proposed an approach based on social relations and content to predict the user's current location.[6] Rui Li and his colleagues developed a unified discriminative influence model to profile users' residence locations based on both user-generated contents and social relations.[7] They integrated signals from both tweets and friends in a unified probabilistic framework to address the problems of sparsity and noise. Based on their model, a multiple location profiling model was also proposed to address the problem of multilocations.[8]

In the proposed methods we describe here, the NLP models are leveraged for the user's location as well as an update's location, whereas the social relationship is only leveraged to estimate the users' residence locations. Our work focuses on residence location, which is defined as the place where most of the user's activities occur.

References

1. Z. Cheng, J. Caverlee, and K. Lee, "You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users," *Proc. 19th ACM Int'l Conf. Information and Knowledge Management* (CIKM), 2010, pp. 759–768.
2. S. Chandra, L. Khan, and F. Bin Muhaya, "Estimating Twitter User Location Using Social Interactions: A Content Based Approach," *Privacy, Security, Risk, and Trust: Proc. IEEE 3rd Int'l Conf. Social Computing* (SocialCom), 2011, pp. 838–843.
3. H. Chang et al., "@ Phillies Tweeting From Philly? Predicting Twitter User Locations with Spatial Word Usage," *Proc. Int'l Conf. Advances in Social Networks Analysis and Mining* (ASONAM), 2012, pp. 111–118.
4. L. Backstrom, E. Sun, and C. Marlow, "Find Me if You Can: Improving Geographical Prediction with Social and Spatial Proximity," *Proc. ACM WWW*, 2010, pp. 61–70.
5. A. Sadilek, H. Kautz, and J.P. Bigham, "Finding Your Friends and Following Them to Where You Are," *Proc. 5th ACM Int'l Conf. Web Search and Data Mining*, 2012, pp. 723–732.
6. E. Cho, S.A Myers, and J. Leskovec, "Friendship and Mobility: User Movement in Location-Based Social Networks," *Proc. 17th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining* (SIGKDD), 2011, pp. 1082–1090.
7. R. Li et al., "Towards Social User Profiling: Unified and Discriminative Influence Model for Inferring Home Locations," *Proc. 18th ACM SIGKDD Int'l Conf. Knowledge Discovery And Data Mining* (SIGKDD), 2012, pp. 1023–1031.
8. R. Li, S. Wang, and K. Chen-Chuan Chang, "Multiple Location Profiling for Users and Relationships from Social Network and Content," *VLDB Endowment*, vol. 5, no. 11, 2012, pp. 1603–1614.

inference in social media by jointly considering social, visual, and textual information. First, we propose a data-driven approach to explore the use of friendship locality, social proximity, and content proximity for geographically nearby users. Based on these observations, we then propose a location propagation algorithm to effectively infer residence location for social media users. We extensively evaluate the proposed method using a large-scale real dataset, and the results demonstrate that we can achieve 15 percent relative improvement over state-of-the-art approaches.

Locations in Social Media

Locations help us find users or tweets in social media. Users provide pieces of location information in their content and profiles, such as text, images, tags, profiles, time zones, and GPS coordinates.

The textual content in social media such as Twitter or Weibo can be no longer than 140 characters, and it is often written in unstructured language. Identifying location information from such textual content is difficult because some toponyms are seldom used and cyberwords or abbreviations are widespread. Even the specific place names users include in posts do not always identify their current location. Links in content might include check-in geotags from location-based services[4] or geotagged pictures on Flickr. GPS coordinates are mostly provided by mobile devices and usually identify the user's current location.

Most social media users maintain a profile page. Users can include their residence locations in the location field in their profiles. Still, the location entries in social media profiles are heterogeneous and have a relatively large geographic scope. The location field may also contain fake location information. One study found that only 66 percent of submitted content refers to geographic information and approximately 2.6 percent of users post multiple locations.[5] In addition to the location field, other parts of profiles that could help us infer user locations include the time zone and UTC24-Offset.

Problem Statement

To ease our further description, this section defines the terminology and describes the problem we use here. In a social media platform such as Tencent Weibo (http://t.qq.com), for any given user, we can detect user locations and following relationships between the users. If a user v_i follows v_j, that does not necessary indicate that v_j follows v_i. However, if v_j and v_i follow each other, we define the relationship between v_i and v_j as friends.

We summarize a social media as a graph $G = G(V, E)$, where V is the user set of v_i and E is the relationship set of $e(v_i, v_j)$ from v_i to v_j. Generally, every user v_i is related to a location ℓ_i. We view ℓ_i as a coordinate point (longitude, latitude) in the geographic space. Our goal is to predict the missing locations. Located users are denoted by $V_l = \{v_1, \ldots, v_l\}$, and the unlocated users are denoted by $V_u = \{v_{l+1}, \ldots, v_{l+u}\}$. Using this notation, we can define the problem of user location prediction as follows: given a social graph $G(V, E)$, predict the residence location of each unlocated user $\{v \in V_u\}$ so that the predicted location ℓ_v is close to the true location ℓ_v^{true}.

Location Propagation

We studied approximately 200,000 users sampled from Tencent Weibo (http://t.qq.com), including their IDs, followers, followees, and residence locations. We collected one month of tweets generated or shared by these sampled users. We observed three phenomena: friendship locality, social proximity, and content proximity.

The first observation, *friendship locality,* comes from this hypothesis: users' online social graphs somehow reflect their offline social relations. Considering the spatial limitations in the physical world, we assume that geographically nearby users are more likely to establish friendships, and we hoped to validate this hypothesis using the real data. Figure 1 gives a log-log plot for the probability of friendship versus distance between users. We can observe an obvious power law in this figure that demonstrates the existence of friendship locality. Based on this observation, we see that residence location can be propagated along friendship relations, which fundamentally motivated us to adopt a label-propagation model for residence location inference.

Based on the first observation, we further dug into the probability of two users sharing the same residence location: *social proximity.* Given friendship locality, we also assume that geographically nearby users tend to have more common friends. We validated this hypothesis with real data. Figure 2 plots the probability of two users sharing the same residence location versus their percentage of common friends. We can see that the probability of two users sharing the same residence location monotonically increases with the increase of the percentage of their common friends, demonstrating the existence of social proximity. Hence, we can see that social proximity is quantified by the concept that the number of common friends is a key factor of location propagation probability between any pair of users.

The third observation, *content proximity,* deals with user-generated content. Because user-generated content in social media mainly comes from users' lives in the physical world, geographically nearby users may tend to publish similar georelated content. To validate this assumption, in Figure 3, we plotted the probability of two users sharing the same residence location versus the similarity of their generated content. We can see that the probability of two users sharing the same residence location monotonically increases with the increase of content similarity in both modalities. Thus, the content proximity is another key factor of location propagation probability between any pairwise users.

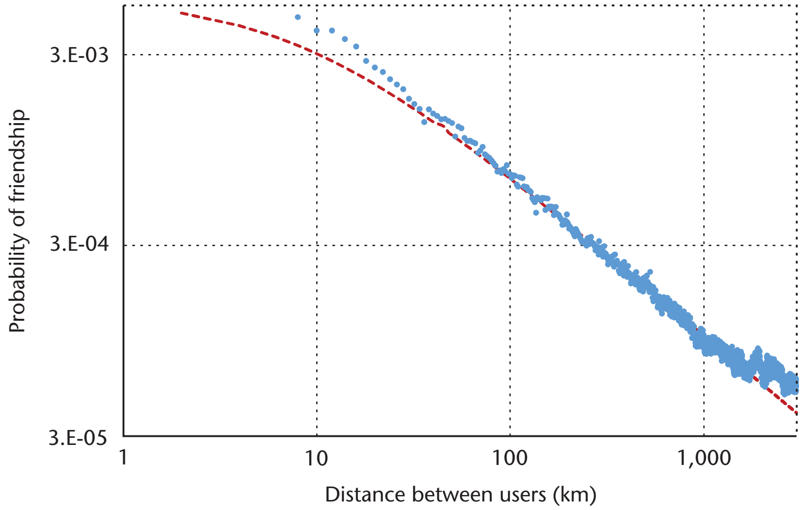


Figure 1. Friendship locality. The Tencent Weibo user data shows that residence location can be propagated along friendship relations.

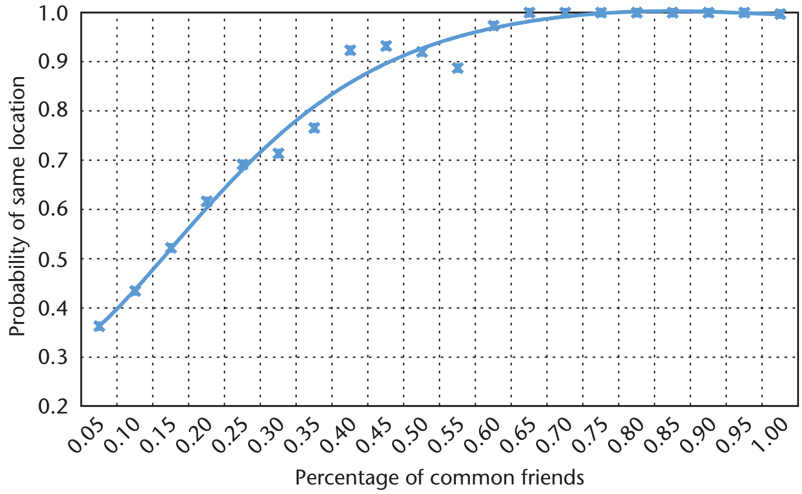


Figure 2. Social proximity. The probability of two users sharing the same residence location increases with the percentage of common friends.

Location Propagation Probability

Based on all these observations and insights, we propose a social-content joint location propagation framework. From the raw data of social relations and profiles as well as textual and visual user-generated content, we extract a social graph with known locations as the propagation medium, calculate the content and social proximity to define the propagation probability, and integrate them into a location propagation algorithm to infer the residence locations for unknown users.

Definition 1 (Content Proximity). The content proximity is defined as a linear combination of textual content similarity and visual content similarity:

$$\mathrm{P}_{\mathrm{con}}(i,j) = \beta S_{\mathrm{txt}}(i,j) + (1-\beta) S_{\mathrm{vis}}(i,j),$$
$$0 < \beta < 1.$$

Here we represent the georelated textual content using word vectors and georelated images using visual word vectors. Then we use the cosine distance to measure the content proximity.

Definition 2 (Social Proximity). We define the social proximity as the Jaccard distance between the target users:

$$\mathrm{P}_{\mathrm{con}}(i,j) = \frac{|F_i \cap F_j|}{|F_i \cup F_j|}.$$

Definition 3 (User Similarity). The similarity P_{ij} is defined as a linear combination of social proximity $\mathrm{P}_{\mathrm{con}}(i,j)$ and content proximity $\mathrm{P}_{\mathrm{con}}(i,j)$:

$$\mathrm{P}_{ij} = \alpha \mathrm{P}_{\mathrm{con}}(i,j) + (1-\alpha)\mathrm{P}_{\mathrm{soc}}(i,j), 0 < \alpha < 1.$$

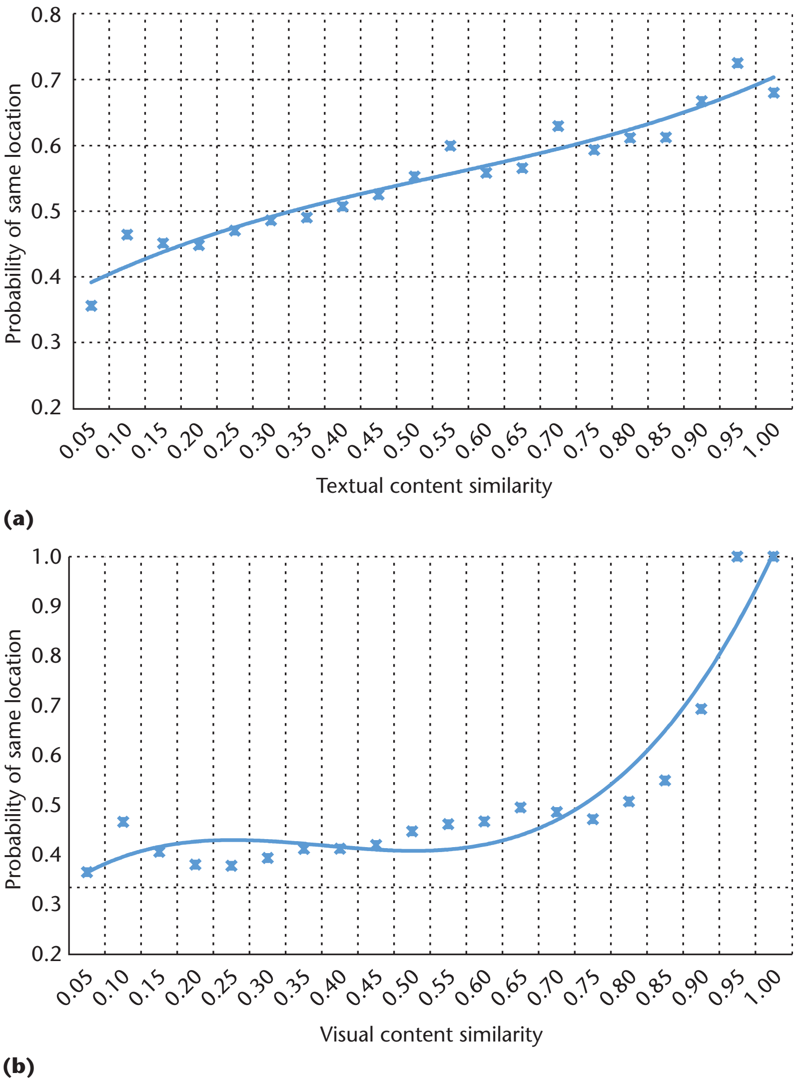


Figure 3. Content proximity. The probability of two users sharing the same residence location increases with their (a) textual content and (b) visual content similarity.

Definition 4 (Location Propagation Probability). The location propagation probability $t_{i,j}$ denotes the probability of location propagation from user v_i to user v_j:

$$t_{i,j} = P(i \rightarrow j) = \frac{w_{ij}}{\sum_{k=1}^{l=u} w_{kj}}.$$

We use a standard normal distribution to calculate the weight w_{ij}:

$$w_{ij} = \exp\left\{-\frac{\mathrm{P}_{ij}^2}{2\sigma^2}\right\}.$$

Model Formulation

In this article, we extend the label propagation algorithm used in previous work[6] to predict locations. Algorithm 1 (see Figure 4) is a semi-supervised, iterative algorithm designed to infer labels for items connected in a network. Usually, the true labels are known for only a small fraction of network nodes, which serve as a source of ground truth to infer the labels of the other nodes. The algorithm proceeds iteratively, and in each round, items receive the label that occurs most frequently for their neighbors. Herein, we apply the label term to the residence location term.

We define a $(l+u) \times (l+u)$ probability propagation matrix $\mathbf{T}$ to measure the probability of location propagation from users to their friends:

$$\mathbf{T} = \begin{Bmatrix} t_{1,1} & t_{1,2} & \dots & t_{1,l+u} \\ t_{2,1} & t_{2,2} & \dots & t_{2,l+u} \\ \vdots & \vdots & \ddots & \vdots \\ t_{l+u,1} & t_{l+u,2} & \dots & t_{l+u,l+u} \end{Bmatrix}.$$

We define the location label normalized matrix $\mathrm{L}_{(l+u)\times p}$, where p is the number of locations. Initialize the matrix as

$$\mathrm{L}_{ij}^{(0)} = \begin{cases} 1; & \ell_{ij} \text{ is the location of user } v_i \\ 0; & \text{else} \end{cases}.$$

We theorize that the location propagation prediction algorithm converges:

$$\begin{aligned} \mathrm{L}^{(t)} - \mathrm{L}^{(t-1)} &= \mathbf{T}^{(t)}\mathrm{L}^{(0)} - \mathbf{T}^{(t-1)}\mathrm{L}^{(0)} \\ &= \mathbf{T}^{(t-1)}[\mathbf{T} - 1]\mathrm{L}^{(0)}. \end{aligned}$$

Every row of the probability propagation matrix $\mathbf{T}$ is nonnegative and the sum is 1, so

$$\lim_{t\to\infty} \mathbf{T}^{(t-1)} = 0.$$

Hence L must converge, so the algorithm converges.

Experimental Results

We first sampled 2 million users from the complete dataset of the Tencent microblog, including their IDs, followers, followees, and residence locations. We also collected one month of microblogs generated or shared by these sampled users. For the residence locations, we extracted both city-level (city, province) and province-level locations from their profiles. In total, our dataset includes 355 cities and 31 provinces as the pool of residence locations.

By identifying city and province names from the text in location profiles, we obtained the corresponding latitude and longitude pairs, calculated the geodistance between cities of users, and observed the relationship between probability of friendship. The curves in Figure 1 show that the probability distribution exhibits a power law.

We first identified users with city-level locations and then randomly selected 500,000 users from them to produce our testbed. Among them, we randomly selected 100,000 located users with at least 20 microblogs and 20 labeled followers and followees.

Algorithm 1

Require: $G = (V, E)$; $\{v_1, \ell_1\}, \ldots, \{v_l, \ell_1\}$

Ensure: $\{v_l + 1, \ell_1 + 1\}, \ldots, \{v_l + u, \ell_1 + u\}$

1: Calculate weight of user similarity w_{ij}

2: Calculate the propagation matrix **T**

3: Initialize $L^{(0)}$

// The u bottom lows are assigned as 0

4: **for** $t = 0$; $L^{(t)}$ converges; $t++$ **do**

5: $L^{(t)} = \mathbf{T}\, L^{(t-1)}$

// In the tth iteration, each user receives the location propagated of friends according to the similarity matrix, updates its probability distribution

6: Clamp the labeled data

// Keep the initial locations of located users

7: **end for**

8: Return location of unlocated user.

Figure 4. Location propagation algorithm. Algorithm 1 is a semisupervised, iterative algorithm designed to infer labels for items connected in a network.

Evaluation Methods

We compared our method with the state-of-the-art methods based on social graphs.[6,7] A social network based approach (FDM) was proposed to predict a user's location based on a social graph that treats all followers and followees as the user's friends.[6] A unified discriminative influence (UDI) model infers a user's residence locations based on both social and content.[7] The location prediction approach (LPA) is based on our location propagation algorithm.

For our evaluation, we retained 20 percent of the known users as ground truth, used the difference between predicted location and the ground truth location to evaluate the performance, compared the accuracy of different approaches both at the city and province levels, and showed the effectiveness of location propagation by comparing it with the baseline approaches.

For each test user $v_i \in V$, we calculated the error distance (Err_i), which represents the distance between the predicted location ℓ_i and the true residence location ℓ_i^{true}:

$$\mathrm{Err}_i(v_i) = \mathrm{EarthDist}\left(\ell_i, \ell_i^{\mathrm{true}}\right).$$

We defined the average error distance (AED) and accuracy (ACC) as

$$\mathrm{AED} = \frac{\sum_{v_i \in V} \mathrm{Err}(v_i)}{|V|}$$

and

$$\mathrm{ACC} = \frac{\left|V \cap \{v_i | \ell_i = \ell_i^{\mathrm{true}}\}\right|}{|V|}.$$

*Table 1. Accuracy comparison table.**

Evaluation method	FDM	UDI	LPA
ACC city (%)	52.1	59.4	68.2
ACC province (%)	58.3	71.5	73.7
AED 80 percent	320	251	238
AED 100 percent	987	830	783

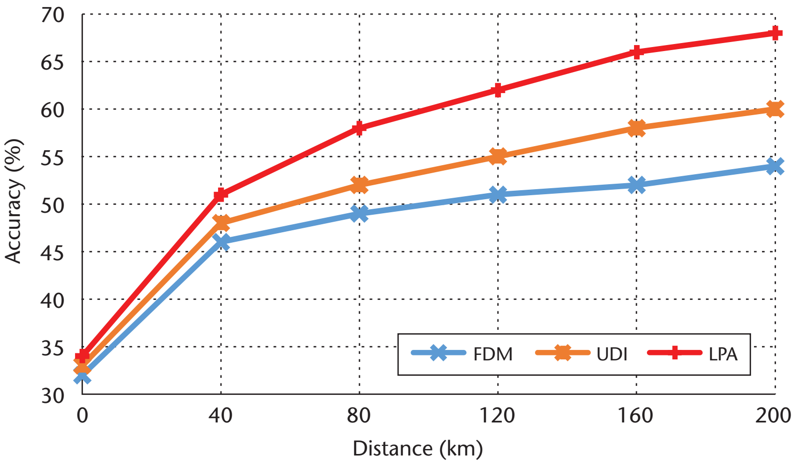
*Social network based approach (FDM), unified discriminative influence (UDI) model, and location prediction approach (LPA).

Results

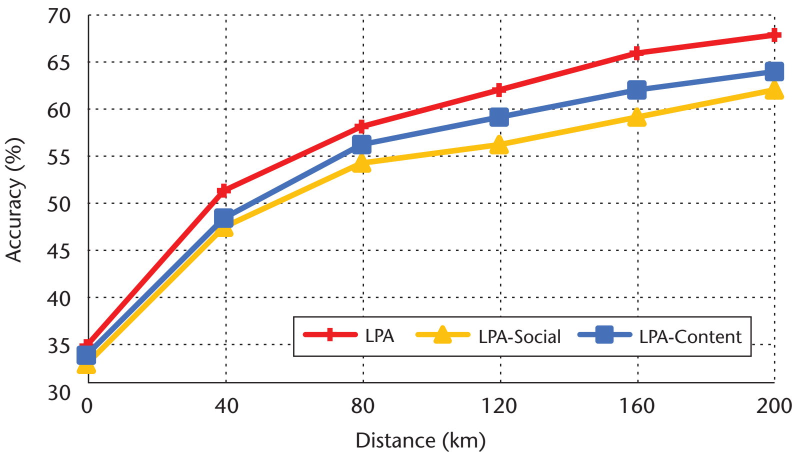
Table 1 shows the performance of our method as well as the FDM and UDI approaches, all three of which profile user locations based on social graphs. The results show that our approach outperforms the baseline approaches.

The AED results in Table 1 show an improvement over the baseline approaches. Because AED is easily influenced by outliers, we report AED at different percentages. AED at x percent denotes that the average error distance of the top x percent of predictions. When we compare AED 80 percent and AED 100 percent, the average error distance increases to 800 km rapidly because the average error distance is influenced by the users predicted inaccurately. Hence, we do not just pay attention to AED 100 percent.

Table 1 also shows that the location propagation algorithm yielded promising accuracy. LPA improves in accuracy by 15 percent at the city level and 20 percent at the province level over the baseline approaches.



(a)



(b)

Figure 5. Accumulative accuracy at various distance. (a) LPA versus baselines and (b) LPA trade-off.

Table 2. Accuracy trade-off table.

Evaluation method	LPA social	LPA content	LPA
ACC city (%)	62.5	64.2	68.2
ACC province (%)	67.1	69.3	73.7
AED 80 percent	250	247	238
AED 100 percent	810	800	783

To explore our experimental results more closely, we plotted accumulative accuracy curves at various distances for each approach (see Figure 5a). A coordinate point (x, y) in the curve shows that y percent of users are correct for x kilometers. The curves show that LPA is more accurate than the baseline approaches at different distances. This is because the LPA approach assumes that more friends are close by, while the baseline approaches require that nonfriends be further away, which may not always be true.

To get more insight into the proposed method, we implemented two variants. One is a purely social-based LPA, and the other is a content-based LPA. Table 2 and Figure 5b give the experimental results, which demonstrate that both social proximity and content proximity contribute significantly to residence location inference.

Finally, we compared the efficiency of our approach and the baseline approaches. Our approach is almost constant, while the baseline approaches are linear. When the number of users is low, our approach and the two baseline approaches take approximately 2 seconds. As the number of users increases, however, the running time of our approach remains almost constant, whereas that of the baseline approaches increases rapidly. Our approach only considers a user's friends, which is an almost constant number. The baseline approaches, on the other hand, consider all users, including both friends and nonfriends, so it is linearly correlated to the volume of the dataset. Thus, our approach is much more efficient and scalable.

Conclusion

Location inference for social media users is a vital problem, and we could further extend our current methods by combining the temporal feature. It can help us study users' migration over time. In addition, we can extend our approach for other attributes, such as users' gender or occupation. **MM**

Acknowledgments

This work is supported by the Natural Science Foundation of China (grants 60933013 and 61003097), 973 Program (grant 2011CB302206), National Significant Science and Technology Projects of China (grant 2011ZX01042–001–002), and Tsinghua-Tencent Joint Lab for Internet Innovation Technologies.

References

1. T. Tumasjan et al., "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," *Proc. 4th Int'l AAAI Conf. Weblogs and Social Media* (ICWSM), 2010, pp. 178–185.
2. E. Aramaki, "Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter," *Proc. Conf.*

Empirical Methods in Natural Language Processing (CEMNLP), 2011, pp. 1568–1576.
3. K. Starbird et al., "Chatter on the Red: What Hazards Threat Reveals about the Social Life of Microblogged Information," *Proc. ACM Conf. Computer Supported Cooperative Work* (CSCW), 2010, pp. 241C–250.
4. C. Cheng et al., "Fused Matrix Factorization with Geographical and Social Influence in Location-Based Social Networks," *Proc. Assoc. Advancement of Artificial Intelligence* (AAAI), 2012, vol. 12, p. 1.
5. B. Hecht et al., "Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles," *Proc. SIGCHI Conf. Human Factors in Computing Systems* (CHI), 2011, pp. 237–246.
6. X. Zhu and Z. Ghahramani, "Learning from Labeled and Unlabeled Data with Label Propagation," tech. report CMU-CALD-02-107, Carnegie Mellon Univ., 2002.
7. L. Backstrom, E. Sun, and C. Marlow, "Find Me if You Can: Improving Geographical Prediction with Social and Spatial Proximity," *Proc. ACM WWW*, 2010, pp. 61–70.
8. R. Li et al., "Towards Social User Profiling: Unified and Discriminative Influence Model for Inferring Home Locations," *Proc. 18th ACM SIGKDD Int'l Conf. Knowledge Discovery And Data Mining* (SIGKDD), 2012, pp. 1023–1031.

Dan Xu is a PhD candidate in the Department of Computer Science and Technology at Tsinghua University, Beijing. His research interests include online social network, security, and privacy. Xu has an ME in mathematics from the University of Information Engineering. Contact him at xud09@mails.tsinghua.edu.cn.

Peng Cui is an assistant professor in the Department of Computer Science and Technology at Tsinghua University, Beijing. His research interests include data mining, multimedia processing, and social network analysis. Cui has a PhD in computer science from Tsinghua University. He is a member of IEEE. Contact him at cuip@tsinghua.edu.cn.

Wenwu Zhu is a professor in the Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology at Tsinghua University, Beijing. His research interests include wireless and Internet multimedia communication and computing. Zhu has a PhD from the Polytechnic Institute of New York University. He is a fellow of IEEE. Contact him at wwzhu@tsinghua.edu.cn.

Shiqiang Yang is a professor in the Department of Computer Science and Technology at Tsinghua University, Beijing. His research interests include multimedia procession, media streaming, and online social network. Yang has an ME in computer science from Tsinghua University. He is a senior member of IEEE. Contact him at yangshq@tsinghua.edu.cn.