**Noname manuscript No.**
(will be inserted by the editor)

# SNS User Classification and its Application to Obscure POI Discovery

**Chenyi Zhuang · Qiang Ma ·**
**Masatoshi Yoshikawa**

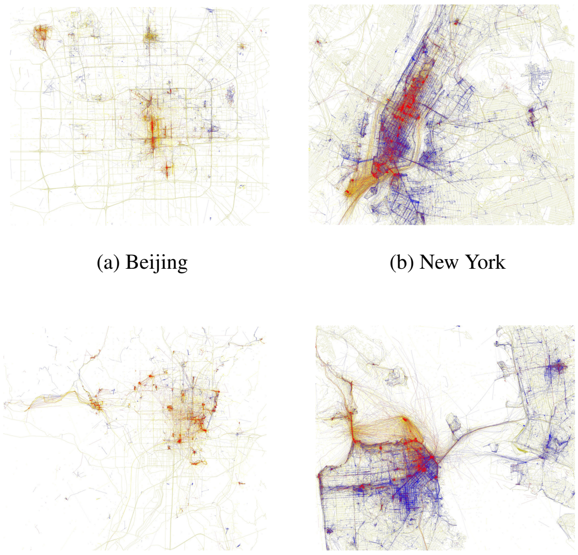
**Abstract** Technologies are increasingly taking advantage of the explosion of social media (e.g., web searches, ad targeting, personalized geo-social recommendations, urban computing). Estimating the characteristics of users, or user profiling, is one of the key challenges for such technologies. This paper focuses on the important problem of automatically estimating *social networking service (SNS) user authority* with a given city, which can significantly improve location-based services and systems. The “authority” in our work measures a user’s familiarity with a particular city. By analyzing users’ social, temporal, and spatial behavior, we respectively propose and compare three models for user authority: a social-network-driven model, time-driven model, and location-driven model. Furthermore, we discuss the integration of these three models. Finally, by using these user-profiling models, we propose a new application for geo-social recommendations. In contrast to related studies, which focus on popular and famous points of interests (POIs), our models help discover obscure POIs that are not well known. Experimental evaluations and analysis on a real dataset collected from three cities demonstrate the performance of the proposed user-profiling models. To verify the effect of discovering obscure POIs, the proposed application was implemented to discover and explore obscure POIs in Kyoto, Japan.

**Keywords** User Profiling · Obscure Points of Interests · Probabilistic Model · Social Network

## 1 Introduction

The explosion of social networking services (SNSs), such as Twitter, Facebook, and Flickr, has led to a wealth of research into using social media content and various social graphs. For example, major search engines, such as Google, Bing, and Yahoo, now incorporate user-generated content (UGC) and trend analysis in their results. Moreover, in order to meet the diverse requirements of users, a significant effort has been dedicated to developing new applications.

Chenyi Zhuang · Qiang Ma · Masatoshi Yoshikawa
Department of Social Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku,
Kyoto, 606-8501, JAPAN
E-mail: zhuang@db.soc.i.kyoto-u.ac.jp, {qiang, yoshikawa}@i.kyoto-u.ac.jp



(a) Beijing (b) New York

(c) Kyoto (d) San Francisco

**Fig. 1** Mobility behavior of different user groups. Red: tourists; blue: locals; yellow: either. [10]

In this context, one problem of significant interest is that of *location-authority-based user profiling*—i.e., mining the information asymmetry regarding a location among various online user groups. That is, the "location authority" measures a user's familiarity with a particular location. In our work, by analyzing different kinds of user behavior that are the results of information asymmetry, location authority can be measured. These vital user models can help in different application scenarios, including the following:

**Geo-social recommendations**: according to a survey by Zheng et al. [1], the growing geo-referenced and community-contributed media resources have generated considerable amounts of detailed location and event tags, covering not only famous landmarks but also obscure locations. Although obscure locations are less well known, they are often still worth visiting, and they are a valuable sightseeing resource for development and promotion. Recently, several studies [2–5] have shown the potential of UGC mining for discovering popular or famous locations and landmarks. Nevertheless, this area of research is relatively unexplored. Because the existence of obscure locations is discovered from the information asymmetry among different user groups, users' location-authority information can be used to discover them. For example, some beautiful locations may be popular with residents but relatively unknown to tourists who are unfamiliar with the city.

**User profiling**: a number of algorithms have been proposed for detecting the different attributes of users. For instance, conventional methods for location inference have been proposed from a wide variety of social media, such as Tweets on Twitter [6], geo-referenced pages on Wikipedia [7], image tags on Flickr [8], and social-network structures on Facebook [9]. Unlike these related methods, location authority focuses on estimating the authority of a user regarding some particular topic.

An intuitive example of this is shown in Figure 1. By analyzing the time information in images, Fischer [10] presented the city-wide mobility behavior of Flickr users. As seen in these four cities, locals have significantly different mobility behaviors when taking photographs compared to tourists. Further analysis shows that the reason for this is the information asymmetry between these user groups. In this example, those familiar with the city (i.e., locals) have more choices for sightseeing (e.g., enjoying cherry blossom). That is, they are the authoritative users of the target city. By analyzing image content, we can further detect the topics with which they are familiar. However, visitors are sometimes more authoritative than residents, and vice versa. For example, some locations are better known to professional photographers than they are to residents. Accordingly, location-authority analysis cannot simply depend on the location information specified in online user profiles. Moreover, according to our statistics, most users do not disclose information regarding their residency.

Therefore, in this paper, we mainly focus on the location authority based user-classification task. By analyzing geo-tagged UGC, three complementary models are proposed. First, we introduce a social-network-driven model, where a graph-based method is used to determine whether a user is living in the target city. Second, we introduce a time-driven model, which estimates the user's location authority based on the frequency of visits. Third, considering the fact that different users have different mobility behaviors, a location-driven model is devised. Furthermore, in response to the shortages of each model, two ensemble methods are discussed. Finally, utilizing these methods, an application is implemented to discover points of interest (POIs) that are not well known, even though they are worth visiting. To the best of our knowledge, this is the first attempt at discovering such POIs.

In summary, we make the following major contributions:

- From a theoretical standpoint, we devise a comprehensive solution to classify SNS users on the basis of their location authority. Thus, the influence of the social network and the temporal and spatial influence from geo-tagged UGC are analyzed. We further propose two ensembles of these models, in which each user's available information is dynamically considered in order to choose models with higher discriminative ability.
- From an application standpoint, by utilizing our proposed models, we propose a new application to discover obscure POIs. To the best of our knowledge, this study is among the first of its kind to discover these little known locations.
- We evaluate the proposed methods with experiments on a real-world dataset collected from three cities on Flickr. By implementing an application, we demonstrate the discovery of obscure POIs using these models.

The remainder of the paper is organized as follows: Section 2 places our research in the context of previous related work; Section 3 presents the proposed models and their ensemble methods; Section 4 outlines our experiments and presents an in-depth quantitative and qualitative analysis of the obtained results; Section 5 describes an application of our models; Section 6 concludes our work.

## 2 Related Work

In this section, we discuss some research related to our study, including user-attribute detection and location discovery.

**User-attribute Detection:** A number of algorithms have been proposed for estimating the home location of Twitter users using content analysis. For instance, Eisenstein et al. [11] built geographic topic models to predict the locations of users in terms of regions and states. Using a Gaussian Mixture Model and Maximum Likelihood Estimation, Chang et al. [6] proposed a city-level location-estimation method. Estimating at city level is more challenging than estimations at higher granularities, such as states or countries. In addition to the analysis of Tweets, other methods for inferring location have been proposed based on social media, including geo-referenced pages on Wikipedia [7], image tags on Flickr [8], and social network structures on Facebook [9].

Recently, several studies have focused on predicting the places where users will go next, based on their previously generated content. For instance, Gao et al. [12] explored the patterns of user check-ins and built a predictive model for user check-in behavior. They found that friends of users tend to go to similar locations, and that the users' visits follow a power-law distribution. This means that users tend to visit only a few places multiple times and many places only a few times. Cho et al. [13] developed a periodic and social mobility model. Their model is not designed to detect the home location of Twitter users, but rather their mobility patterns—e.g., when the user is at "home" and when the user is at "work."

In addition to location information, a variety of other user attributes have been investigated using natural language processing, including gender, age, and even political orientation [14, 15]. The problem of automatically constructing user profiles has been commonly regarded as important. Nevertheless, to the best of our knowledge, estimating a user's location authority by analyzing online generated social media has not yet been thoroughly investigated.

**Location Discovery:** In an effort to discover POIs, a survey given by Luo et al. [16] shows that collections of geo-multimedia data, which are a result of sightseeing experiences shared in web communities, are widely used in trip recommendations. Ji et al. [4] modeled the relationships of scene/landmark and scene/authorship as a graph and adopted two popular link-analysis methods, PageRank and HITS, to mine representative landmarks. Zhang et al. [17] aimed to mine interesting locations and classical travel sequences in a given geospatial region on the basis of multiple users' GPS trajectories. They first modeled multiple individuals' location histories with a tree-based hierarchical graph. Then, using the graph, they proposed a HITS-based inference model that infers the interest of a given location. In [18], the authors further developed a recommendation system. Instead of GPS traces, Liu et al. [5] proposed a joint authority analysis framework to discover areas of interest with geo-tagged images and check-ins. Hasegawa et al. [19] attempted to organize travel-related Tweets by considering the spatio-temporal continuity of user behavior during travel. By merging these fragmented Tweets, users' travel experiences can be detected.

In these studies, GPS traces, images, check-ins, and Tweets are treated as different kinds of user votes to help gather tourism knowledge. Conventional methodologies, such as "rank-by-count" and "rank-by-frequency" in a voting manner, are the basis for most of these trip-recommendation methods. Owing to the lack of related raw data on the Internet, however, methods for discovering obscure rather than popular locations have not been well investigated.

## 3 General Solution to Users' Location-authority Analysis

In this section, we describe in detail three types of information that can help to characterize a user's location authority: *social network*, *visiting frequency* and *mobility behavior*. In accordance with these three types of information, three models (i.e., social-network-driven, time-driven, and location-driven models) are introduced in order to estimate SNS users' authority with a particular city. Our goal is two-fold: first, to provide an assessment of the robustness and generalization potential of features for authority-based user-classification purposes; and second, to explore methods of combining these three models for higher classification performance. We first describe each individual model before discussing their combination.

### 3.1 Social Network: “Is the user living in the target city?”

For target city $c$, the most direct way of identifying a user’s authority is to retrieve residency information from the user’s profile. However, few users are willing to disclose this information or even their country of residence. In accordance with the statistics from our experimental dataset, 64.8% of the 1,827,500 Flickr users did not specify any information regarding their residency. Moreover, although the other 35.2% users did provide location information, they were not necessarily familiar with the location specified. A reasonable explanation for this exception is that users commonly provide their hometown rather than their current location.

Therefore, we propose a social-network-based method for calculating the authority of a user with given city $c$, by which “authoritative communities” can be detected on a network. The assumption for this model is that a user with many friends who are familiar with city $c$ will also be familiar with $c$. The statistics compiled in [9, 20] confirm that geographic and online social relationships are inextricably intertwined. Generally, the likelihood of friendship is inversely proportional to distance. Unlike the related studies [9, 20] that investigated pairwise friendship between two users using a maximum likelihood approach, our algorithm further aggregates collective relationships by information propagation on the social network. Before introducing the algorithm, we first present the definition of a social network.

**Definition 1** A social network $N$ is a directed graph $N(V,E)$, where $V$ is the user set of $v_i$, and $E$ is the friendship set of $e(v_i,v_j)$ from $v_i$ to $v_j$. In our implementation, the network is represented as a transition matrix:

$$N(v_i,v_j)=\begin{cases} 0 & if(v_j,v_i)\notin E,\\ 1/outdegree(v_j) & if(v_j,v_i)\in E. \end{cases} \tag{1}$$

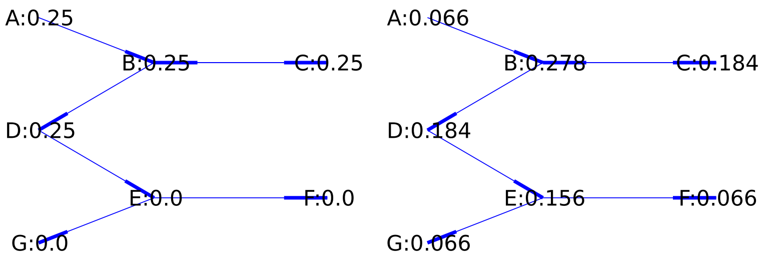
We then devise a biased authority-propagation algorithm, Algorithm 1, to model this. By

**Algorithm 1** Biased Authority-propagation Algorithm.

```
calculate $d_c$                                   ▷ based on whether or not $v_i$ is a local user
$M_B = 0; \alpha_B = 0.85$
$F_s(c) = d_c$                                    ▷ initialize $F_s(c)$
while $M_B < 1000$ and $F_s(c)$ not converged do
    $F_s(c) = \alpha_B \cdot N \cdot F_s(c) + (1-\alpha_B)d_c$.
    $M_B{+}{+}$
end while
```

iterating the biased algorithm on network $N$, the authority value of a parent node is split among its children and the authority value of a child node is the sum of the authority values propagated over its links. This is exactly consistent with our assumption.

For the normalized bias distribution $d_c$ in Algorithm 1, in contrast to TrustRank [21], our method can automatically select good seeds based on whether user $v_i$ is a local of city $c$, as detected from the user’s profile. The entries for vector $d_c$ that correspond to good seeds sum up to 1. Given city $c$, $F_s(c)$ is initialized to $d_c$ before the iteration. Here, $\alpha_B$ is the decay factor. Figure 2 illustrates an example when nodes [A, B, C, D] are good seeds. After the iteration, we find that node B has obtained the highest score because it was recommended by more good seeds.



(a) Before iteration (b) After iteration

**Fig. 2** Social-network-driven method. Each node represents a user. Edges denote a friendship between two users. The scores measure the authority levels.

Because of the exception above (note that a user may not be familiar with the location specified in their profile), some nodes can be wrongly selected as good seeds in $d_c$. Fortunately, our method can significantly reduce this impact. Our iterative algorithm not only aggregates the "authoritative communities" (e.g., [B, C, D, E] in Figure 2), it also filters out the noise in $d_c$ (e.g., [A]).

Although this model is effective in most cases, it cannot cover situations where potentially good seeds are sparsely distributed (e.g., [F, G]). In other words, if there are many users with few friends on the social network, the sensitivity (or "recall") of this model will be low. We discuss this issue in detail when evaluating the proposed methods.

### 3.2 Visiting Frequency: "Does the user visit the target city frequently?"

The second method is a time-driven method. This method considers the frequency of visits made by users. Intuitively, if a user has visited target city $c$ frequently and recently, the user is familiar with $c$. First, for each user $v_i$, we define matrix $M_i$ to record the user's visitation history.

**Definition 2** Matrix $M_i$ is a mapping between user $v_i$ and all the cities $C_i$ that the user has been to. For example, retrieving $v_i$'s geo-tagged images, $M_i[y = 2014, c = kyoto] = 10, c \in C_i$ means that among $v_i$'s images taken in Kyoto in 2014, there are 10 different dates detected.

On the basis of matrix $M_i$, we calculate the user's authority with each city $c \in C_i$. To do so, three factors must be considered:

1. For each year, the proportion of days each $c$ accounts for is recorded in $M_i$;
2. The revisitation to cities is denoted as matrix $r_i$. Assuming that we collect 10 years' worth of raw data for $v_i$, for each year $x$, $r_i[x,c] = \frac{5}{10}$ when we detect that $v_i$ went to $c$ during five of those years;
3. The staleness of information. To identify whether $v_i$ has been to these cities recently, we introduce diagonal matrix $\omega_i(x,y)\ \forall x,y \in \{1,2,...,n\}$, to characterize this feature. Here, $n$ denotes the number of detected years.

The time-driven method is then used to compute authority score $F_t$ as follows:

$$F_t(v_i,c) = diag(r_i^T \cdot (w_i \cdot M_i))[c], \tag{2}$$

where *diag* denotes the matrix's diagonal vector and $[c]$ is used to obtain the authority score of target city $c$.

Obviously, the accuracy of user $v_i$'s $F_t$ depends to a considerable extent on the sparsity of matrix $M_i$. For a user who has uploaded few geo-tagged UGC (e.g., Tweets, images, check-ins), this method will be ineffective.

## 3.3 Mobility Behavior: "Where do users visit in a target city?"

In addition to the social network and time information introduced above, the geographic information regarding where a user has been in city $c$ is also essential for estimating location authority. Furthermore, unlike our previous two models and related approaches [6] [9] [13], which focus on individual behavior, this third model simulates collective mobility behavior. By introducing latent variable *group* $g$ this model provides a more general solution to simulate the mobility behavior of different user group $g$s. In the next part, we first describe the modeling process followed by the model's parameter inference.

### *3.3.1 Mobility Behavior Modeling*

Table 1 shows the notations used for this model. To generate the map shown in Figure 1, Algorithm 2 uses a generative process. By introducing Bayes' theorem, our algorithm models users' mobility behavior based on the following observations.

1. **Observation 1:** For each user $v_i$, given city $c$, there is an inherent property that describes their authority with the target city. For instance, there are three different groups, labeled red, blue, and yellow in Figure 1, which represent the different mobility behavior of these groups. Accordingly, with our location-driven model, we introduce latent variable *group*, denoted as $g$, to represent this property.
2. **Observation 2:** A region $r$ (see Definition 3, below) is visited by groups $g$ with mixture distributions. That is, users from different groups prefer visiting different regions of a city. In Figure 1, it is clear that some regions appeal to more users from the *red group*, while others attract many users from the *blue group*. In Algorithm 2, this observation corresponds to the generative process from lines 5 to 12.
3. **Observation 3:** A group $g$ is a mixture distribution of users $v$. Similar to the concept of a fuzzy set, given group $g$, each $v_i$ has a probability that indicates their degree of membership to $g$. In Algorithm 2, the generative process from lines 2 to 4 represents this observation.

**Definition 3** Region $R$ is a collection of regions clustered by Gaussian Mixture Model (GMM), as shown in line 1 of Algorithm 2. GMM is used because it is widely applied to simulate users' mobility behavior [6] [13]. Another, more important reason for applying a GMM is to have a unified parameter inference process in Section 3.3.2 that operates in a probabilistic manner.

The whole generative process in Algorithm 2 can be explained using the following joint distribution:

$$p(t,g \mid \alpha,\beta,\Lambda) = p(l_t \mid r,\Lambda)p(v_t \mid g,\alpha)p(g \mid \beta). \tag{3}$$

Given priors $\alpha,\beta$, and $\Lambda$, by introducing latent variable *group* ($g$), the model aims to generate each tuple $t = \{v_t, l_t\}$ (e.g., $t$ represents an image, Tweet, or post) in city $c$. Here,

**Table 1** Notations used for the location-driven model.

| | Size | Description |
|---|---|---|
| $V$ | $[1,V]$ | User set, where $v$ is a particular user in $V$ |
| $R$ | $[1,R]$ | Region set, where $r$ is a region in $R$ |
| $G$ | $[1,G]$ | *Group* set, where $g$ is a group in $G$ |
| $T$ | $[1,T]$ | $t \in T$ is a user-coordinate tuple: $\{v_t, l_t\}$ |
| $\mu_r$ | $\mathbb{R}^2$ | Mean location of latent region $r$ |
| $\Sigma_r$ | $\mathbb{R}^{2\times 2}$ | Covariance matrix of latent region $r$ |
| $\vartheta_{\mathbf{r}}$ | $1 \times \lvert G\rvert$ | Region-dependent group distribution |
| $\varphi_{\mathbf{g}}$ | $1 \times \lvert V\rvert$ | Group-dependent user distribution |
| $\alpha, \beta$ | | Hyperparameters of Dirichlet priors |
| $\Omega$ | $\lvert R\rvert \times \lvert G\rvert$ | $\Omega_{r,g}$: count of tuples in $r$ that are assigned to $g$ |
| $\Psi$ | $\lvert G\rvert \times \lvert V\rvert$ | $\Psi_{g,v}$: count of times when $v$ is assigned to $g$ |
| $\mathcal{N}$ | | Normal distribution |
| $Dir$ | | Dirichlet distribution |
| $Mult$ | | Multinomial distribution |

**Algorithm 2** Probabilistic generative process.

```
$R \sim \sum_{r=1}^{R} \pi_r \mathcal{N}(x \mid \mu_r, \Sigma_r)$        ▷ initialize $R$ using GMM
for all groups $g \in [1,G]$ do                                        ▷ Observation 3
    sample mixture components $\varphi_{\mathbf{g}} \sim Dir(\alpha)$
end for
for all regions $r \in [1,R]$ do                                       ▷ Observation 2
    sample mixture proportion $\vartheta_{\mathbf{r}} \sim Dir(\beta)$
    for all tuples $t \in [1,N_r]$ in region $r$ do
        sample a coordinate $l_t \sim \mathcal{N}(\mu_r, \Sigma_r)$
        sample a group $g_{rt} \sim Mult(\vartheta_{\mathbf{r}})$
        sample a user $v_{rt} \sim Mult(\varphi_{\mathbf{g_{rt}}})$
    end for
end for
```

$\Lambda$ denotes all the priors of the GMM. More specifically, there is a Dirichlet distribution over the mixing coefficients $\pi_r$, and a Gaussian–Wishart distribution governing the mean $\mu_r$ and precision $\Sigma_r^{-1}$ of each Gaussian component.

### *3.3.2 Parameter Learning*

In accordance with our goal of identifying user authority (i.e., the probability a user belongs to group $g$), the target of model inference is distribution $p(g \mid t)$, where the hyperparameters are omitted. In Eq. 3, when region $r$ is given at the initialization stage, $g$ is independent ($\perp\!\!\!\perp$) of $l_t$. Intuitively, we can regard all tuples $t$ in a specific region $r$ as a *bag of instances of different users* by ignoring the location information. Note that, despite $g \perp\!\!\!\perp l_t \mid v_t, r$, our model is sufficiently flexible to describe a situation where different instances $v_t$ of the same user $v$ can be assigned to different $g$s. Accordingly, we divide the inference procedure into the following two steps.

**Step 1.** $l_t - generation$**:** Because city $c$ is described by a GMM and each region $r$ is a component of the GMM, we utilize a variational Bayesian machinery for the inference of the GMM. As mentioned above, because it is unnecessary to identify the exact location information given $r$, a random coordinate is assigned to each $l_t$ in $r$.

Therefore, we simply utilize the implementation introduced in Section 10.2 of [22]. This offers the advantage that the number of mixture components (i.e., $|R|$) can be automatically identified with a relatively large initial value.

**Step 2.** $v_t - generation$**:** After eliminating the $l_t$, the joint distribution in Eq. 3 is simplified as follows:

$$p(t,g \mid \alpha,\beta) = p(v_t \mid g,\alpha)p(g \mid \beta). \tag{4}$$

Three strategies can be applied to solve our inferential problem: EM with variational inference, EM with expectation propagation, and Gibbs sampling. We selected the Gibbs sampling method because its performance is comparable with the other two, but it is more tolerant to local optimization.

**Algorithm 3** Gibbs sampling algorithm for learning Eq. 4.

```
zero all count matrices $\Omega$, $\Psi$;                                ▷ initialization
for all regions $r \in [1,R]$ do
    for all tuples $t \in [1,N_r]$ in region $r$ do
        sample a group $g_{rt} \sim Mult(1/|G|)$ for $t$
        update region-group count: $\Omega[r,g_{rt}]+1$
        update group-user count: $\Psi[g_{rt},v_t]+1$
    end for
end for
while not finished do                                                   ▷ burn-in and sampling period
    for all regions $r \in [1,R]$ do
        for all tuples $t \in [1,N_r]$ in region $r$ do
            exclude the target $t$: $\Omega[r,g_{rt}]-1$; $\Psi[g_{rt},v_t]-1$
            sample a new group $\tilde{g_{rt}}$, using Eq. 7
            update the counts: $\Omega[r,\tilde{g_{rt}}]+1$; $\Psi[\tilde{g_{rt}},v_t]+1$
        end for
    end for
    if converged and sampling iterations are finished then
        obtain the parameters $\vartheta_{\mathbf{r}}$ and $\varphi_{\mathbf{g}}$, using Eq. 8
    end if
end while
```

We derive the Gibbs updating rule for Algorithm 3, i.e., a conditional distribution for a tuple with index $i=(r,t)$, as follows. Note that $T=\{t_i,T^{\neg i}\}$, $G=\{g_i=x,G^{\neg i}\}$, and the priors $\alpha$ and $\beta$ are omitted. Because $t_i \perp\!\!\!\perp T^{\neg i} \mid G^{\neg i}$, we have:

$$\begin{aligned} p(g_i = x \mid G^{\neg i},T) &= \frac{p(G,T)}{p(G^{\neg i},T)} \\ = \frac{p(T \mid G)}{p(T^{\neg i} \mid G^{\neg i})p(t_i)} \cdot \frac{p(G)}{p(G^{\neg i})} &\propto \frac{p(G,T)}{p(G^{\neg i},T^{\neg i})}. \end{aligned} \tag{5}$$

The joint distribution from Eq. 4 can be derived as:

$$\begin{aligned} p(G,T) &= p(T \mid G,\alpha)p(G \mid \beta) \\ &= \prod_{g \in G}^{G} \frac{B(\Psi_g+\alpha_g)}{B(\alpha_g)} \cdot \prod_{r \in R}^{R} \frac{B(\Omega_r+\beta_r)}{B(\beta_r)}, \\ where\ B(\mathbf{x}) &= \frac{\prod_{k=1}^{dim\,\mathbf{x}} \Gamma(x_k)}{\Gamma(\sum_{k=1}^{dim\,\mathbf{x}} x_k)}. \end{aligned} \tag{6}$$

Here, we first assume symmetric Dirichlet priors, with $\alpha$ and $\beta$ each having a single value. Then, by substituting and simplifying Eq. 5, we obtain the Gibbs updating rule:

$$
\begin{aligned}
&p(g_i = x \mid G^{\neg i}, T) = p(g_i = x \mid G^{\neg i}, t_i = v, T^{\neg i}) \\
\propto\; & \frac{\Psi_{x,v} + \alpha_v - 1}{\sum_{v' \in V}^{V} \Psi_{x,v'} + |V|\,\alpha_{v'} - 1} \cdot \frac{\Omega_{r,x} + \beta_x - 1}{\sum_{x' \in G}^{G} \Omega_{r,x'} + |G|\,\beta_{x'} - 1}.
\end{aligned} \tag{7}
$$

This result is intuitive: the first ratio expresses the probability of $t_i$ under group $g_i$, and the second ratio expresses the probability of group $g_i$ in the region $r$ to which the target sampled $t_i$ belongs.

Finally, with a set of samples from the obtained posterior distribution $p(g \mid t)$ (i.e., the converged $\Omega$ and $\Psi$), the multinomial parameters, $\vartheta_{\mathbf{r}}$ and $\varphi_{\mathbf{g}}$, can be computed as follows:

$$
\begin{aligned}
\varphi_{g,v} &= \frac{\Psi_{g,v} + \alpha_v}{\sum_{v' \in V}^{V} \Psi_{g,v'} + |V|\,\alpha_{v'}}, \\
\vartheta_{r,g} &= \frac{\Omega_{r,g} + \beta_g}{\sum_{g' \in G}^{G} \Omega_{r,g'} + |G|\,\beta_{g'}}.
\end{aligned} \tag{8}
$$

These values correspond to the state of the Markov chain in Gibbs sampling. Insofar as they are conditioned by the past $T$ and $G$, they are used to estimate the predictive distribution over new tuples and groups.

### 3.4 Model Integration

The most important reason for integrating the models is to identify latent groups generated by the location-driven model. Because only the user IDs and geo-coordinates are used, the system cannot automatically know the meanings of the groups when this model is exclusively used (i.e., it is unclear whether the group is familiar or unfamiliar with the city). Within this context, the other two models can be utilized to identify latent groups.

Another reason to integrate the models is that they use distinct kinds of raw data. Integrating them thus makes the classification task much more stable and robust. That is, a more general model is obtained for responding to different situations. To this end, we propose two ensembles: a *support vector machine (SVM)* and *dynamic weighted ensemble (DWE)*. These two ensembles enable our solution to utilize the results obtained from the location-driven model, and they enable us to improve its performance.

***SVM***: With this method, we simply use the scores obtained from the three models as features to train multiple classification models. For instance, if the number of groups $|G| = 2$ in the location-driven model, there will be at most four features for SVM model training. However, because the sum of the two scores from the location-driven model is 1.0, in practice three features are sufficient. In the experiments, combinations of these features, taken two at a time, were tested in order to detect key features.

For the implementation, we utilize SVM with the RBF kernel. The grid-search method introduced in [23] is then applied with three-fold cross-validation to find the best RBF kernel parameter $\gamma$ and penalty parameter $C$ in the SVMs. Finally, the most powerful model is selected by comparing the average classification accuracy.

***DWE***: In machine learning, multiple classifiers are always combined in an ensemble, and this is typically more accurate than using an individual model. The most standard method for doing so is majority voting, where the final classification is the class that receives the most

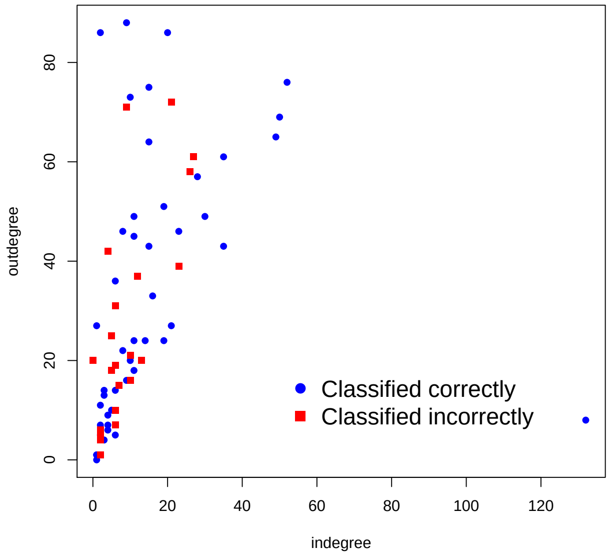


**Fig. 3** Indegree and outdegree distributions of 94 samples from the social network $N$ (see Definition 1). Here, 47 blue samples are correctly identified, while the other 47 red samples are misidentified.

votes from the individual classifiers [24]. In addition, other methods have been proposed, such as bagging [25] and boosting [26].

Before introducing our ensemble strategy, we first discuss the shortcomings of each of the proposed models. For the first social-network-driven model, Figure 3 shows 94 nodes (users) in the friendship-based social network. Among them, 47 blue nodes are correctly classified, while the other 47 red nodes are misclassified. Obviously, there are differences in the available information ("indegree and outdegree" in this case) that can influence the model's performance. Intuitively, for users with few friends (i.e., lower indegree and outdegree), this model would, to a considerable extent, have insufficient discriminative power. The experiment results shown in Figures 6a and 6b further verify this observation. Its performance decreased sharply for the inactive users. Similarly, our experimental results in Figures 6c and 6d show that the second time-driven model's performance decreases sharply for inactive users who generate little content (i.e., images in our case) on the Internet. The performance of the third location-driven model, as shown in Figures 6e and 6f, although not the highest, is the most stable. These observations suggest that we should take the different levels of available information for each user into consideration.

Therefore, in our solution, we use a DWE, the idea for which was first introduced in [27]. In contrast to other methods, this method can account for the differences in the available information for each user. For instance, some users may have more generated data (e.g., images, or Tweets), favoring the time-driven model, whereas other users may have a lot of friends, favoring the social-network-driven model. Moreover, by regarding as evidence users who obtained high weights in the first two models, the DWE can label the latent groups generated by the location-driven model as "familiar with the city" or "unfamiliar with the city."

**Table 2** Datasets collected from the three cities: Beijing, Kyoto, and San Francisco (SF).

| | **Beijing** | **Kyoto** | **SF** |
|---|---|---|---|
| # of geo-tagged images | 4,002 | 6,121 | 3,989 |
| # of target Flickr users | 230 | 300 | 536 |
| % of images from the top 20% | 85.7% | 84.4% | 78.6% |

**Table 3** Contextual data used in Algorithm 1 and Eq. 2.

| | |
|---|---|
| Algorithm 1: | Friendship-based directed social network $N$ (see Definition 1).<br>Nodes: 837,633; Edges: 3,082,408. |
| Eq. 2: | 1,066 matrices $M$ (see Definition 2) recording users' visitation history.<br>Year range: 2005–2015; Cities: 17,460. |

Accordingly, using a user's indegree and outdegree as features, we train learning model $P_s$ to predict the discriminative ability of the social-network-driven model. Similarly, using the amount of UGC as a feature, we train another model $P_t$ for the time-driven model. AdaBoost [26] was used for learning, because it fits a sequence of weak learners. In our case, the learners (i.e., $P_s$ and $P_t$) are decision trees with $max\_depth = 3$.

$$\begin{aligned} F(v_i) &= [F_s, F_t, F_l] \cdot \mathbf{en}^T, \\ where\ \mathbf{en} &= [0.5P_t, 0.5P_s, 1 - 0.5(P_t + P_s)]. \end{aligned} \tag{9}$$

Eq. 9 is used to calculate the assembled authority score of user $v_i$. Here, $F_s$, $F_t$, and $F_l$ are the authority scores or probabilities obtained from the three models, respectively. In order to maintain the same scale, they are first pre-normalized. Ensemble weights **en** are calculated using prediction results of the models $P_s$ and $P_t$. Both $P_s$ and $P_t$ contain two predictions: "1" denotes high classification strength and "0" denotes low classification strength. For example, if neither the social-network-driven model nor the time-driven model have sufficient classification strength, the DWE only uses the stable location-driven model.

In summary, because the three original models are established using completely different information, we expect that their combination will improve the performance. The drawback to such ensembles, however, is that they require a training dataset. Although future research could be directed towards establishing a more comprehensive training dataset, we currently rely on manually labeled random samples. Moreover, in the next section we experimented with cross-validation in order to render the dataset statistically significant.

## 4 Experimental Evaluation

In this section, we outline the retrieved datasets and parameter selections for our experiments. All five user-classification methods (i.e., Algorithm 1, Eq. 2, Eq. 3, SVM, and DWE) are then compared and discussed quantitatively. Finally, some case studies are visualized in order to demonstrate the feasibility of our methods.

### 4.1 Data Preparation

Table 2 summarizes the target datasets for Beijing, Kyoto, and San Francisco, which were collected from Flickr. Unsurprisingly, the 80–20 rule persists on Flickr. That is, for all three datasets, the top 20% Flickr users were prolific, whereas the bottom 80% of the users remained relatively inactive.

Table 3 details the contextual data used by the social-network-driven model (Algorithm 1) and the time-driven model (Eq. 2). We collected a total of 837,633 Flickr users in order to construct social network $N$, in which all the 230 + 300 + 536 target users were included. Furthermore, we retrieved all of the images uploaded by the target users from Flickr. These included all of the geo-tagged images and some other images that were taken just before or after the geo-tagged images. The images were then used to create matrix $M$.

### 4.2 Parameter Selection

To make our models more reproducible, we list all the parameters and their corresponding values.

**Social-network-driven Model:** We applied our biased authority-propagation algorithm by setting the decay factor $\alpha_B = 0.85$, which is often considered to be the default value for PageRank-like calculations in the literature. The authors of [28] further found that when the decay factor changes, the top sections of the ranking changes only slightly. As we are especially interested in "authority communities," i.e. the top sections, the impact of the decay factor's choice is limited. We set the maximum number of iterations $M_B = 1,000$.

**Time-driven Model:** Because the heuristic parameter $w_i$ in Eq. 2 is used to characterize the information staleness, on the basis of its actual effectiveness, we set it to exponential decay: $0.8^{(\lambda-x)}$. That is, for any year $x$, more time passing between $x$ and current year $\lambda$ leads to higher decay.

**Location-driven Model:** We set the number of groups $|G| = 2$. Based on the experimental results below, these two latent groups can be explained as a "tourist-like group" and "local-like group." In accordance with the conclusion in [29], because of the mutual relationship between hyperparameters and the group number, we set the symmetric Dirichlet priors as $\alpha = 0.01$ and $\beta = 50/|G|$. A relatively small value of $\alpha$ can be expected to result in a fine-grained decomposition of the users in region $r$ into groups $g$ that address familiarity. For $\beta$, by keeping constant the sum of the Dirichlet hyperparameters (50 in our case), it can be interpreted as the number of virtual samples contributing to the smoothing of the region-dependent group distribution $\vartheta$. The maximum number of sampling iterations in Algorithm 3 was set to $1,000$.

### 4.3 User-classification Evaluation

**ROC-based Evaluation:** The receiver operating characteristic (ROC) curve is a graphical plot that illustrates the performance of a binary classifier model as its discrimination threshold is varied. In our case, although we attempted to classify users into two groups, we obtained a score or probability for each user rather than a label with two options (i.e., familiar or unfamiliar). Because the determination of an ideal threshold (also known as a cut-off value) is always a tradeoff between sensitivity (true positives) and specificity (true negatives), the ROC curve offers a graphical illustration of these tradeoffs at each threshold.

In our experiments, we used two criteria for the evaluation, as shown in Figure 6: sensitivity (true-positive rate) and 1 - specificity (false-positive rate). Table 4 shows how these these two criteria were calculated.

**Table 4** Calculating the true-positive rate and the false-positive rate.

| **True class** | Familiar (high authority) | Unfamiliar (low authority) |
|---|---|---|
| **Predicted class** | | |
| Familiar | True Positives (TPs) | False Positives (FPs) |
| Unfamiliar | False Negatives (FNs) | True Negatives (TNs) |
| True–positive rate = $TP/P$; <br> False–positive rate = $FP/N$. | | |

**Table 5** Labeling the results of the top 100 active users in each city, denoted by $Active_{set}$.

| **Cities** | **# of familiar users** | **# of unfamiliar users** |
|---|---|---|
| Beijing | $\approx 36$ | $\approx 64$ |
| Kyoto | $\approx 31$ | $\approx 69$ |
| San Francisco | $\approx 76$ | $\approx 24$ |

**Table 6** Labeling the results of the bottom 100 inactive users in each city, denoted by $Inactive_{set}$.

| **Cities** | **# of familiar users** | **# of unfamiliar users** |
|---|---|---|
| Beijing | $\approx 44$ | $\approx 56$ |
| Kyoto | $\approx 15$ | $\approx 85$ |
| San Francisco | $\approx 59$ | $\approx 41$ |

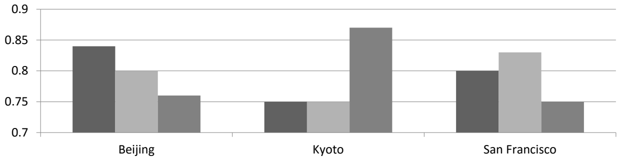


**Fig. 4** Labeling variances on the top 100 active users in each city, calculated using the Jaccard index.

Using these two criteria, the area under the ROC curve (AUC) is recognized as a measurement of a test's discriminatory power. The maximum AUC value is 1.0, indicating 100% sensitivity and 100% specificity, and an AUC value of 0.5 indicates the absence of any discriminative power.

**Ground Truth:** It is almost impossible to acquire the exact ground truth regarding whether a particular Flickr user is familiar with a city. For example, a resident of Beijing is not necessarily familiar with the city. For our experiments, we employed manual efforts to collect an approximate ground truth. This was done by adopting the following procedures.

We invited nine subjects to manually score users' authority over a specific city from the users' online information. For each city (i.e., Beijing, Kyoto, and San Francisco), three independent subjects were assigned to label 200 users that had been selected. In order to obtain a comprehensive evaluation, we selected both users who had uploaded many images to Flickr and those who had not. Table 5 shows the labeling results for 300 active users, from which approximately 36, 31, and 76 familiar users were returned by majority voting. Similarly, Table 6 shows the labeling results for 300 inactive users.

To ensure that the labeling results were credible, all of the subjects were familiar with their corresponding cities, and they recorded the reasons for all of their labels. For each

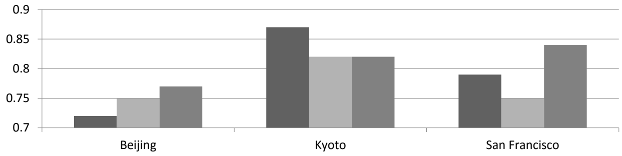


**Fig. 5** Labeling variances on the bottom 100 inactive users in each city, calculated using the Jaccard index.

**Table 7** Classification accuracy of the general SVM models when using different numbers of features.

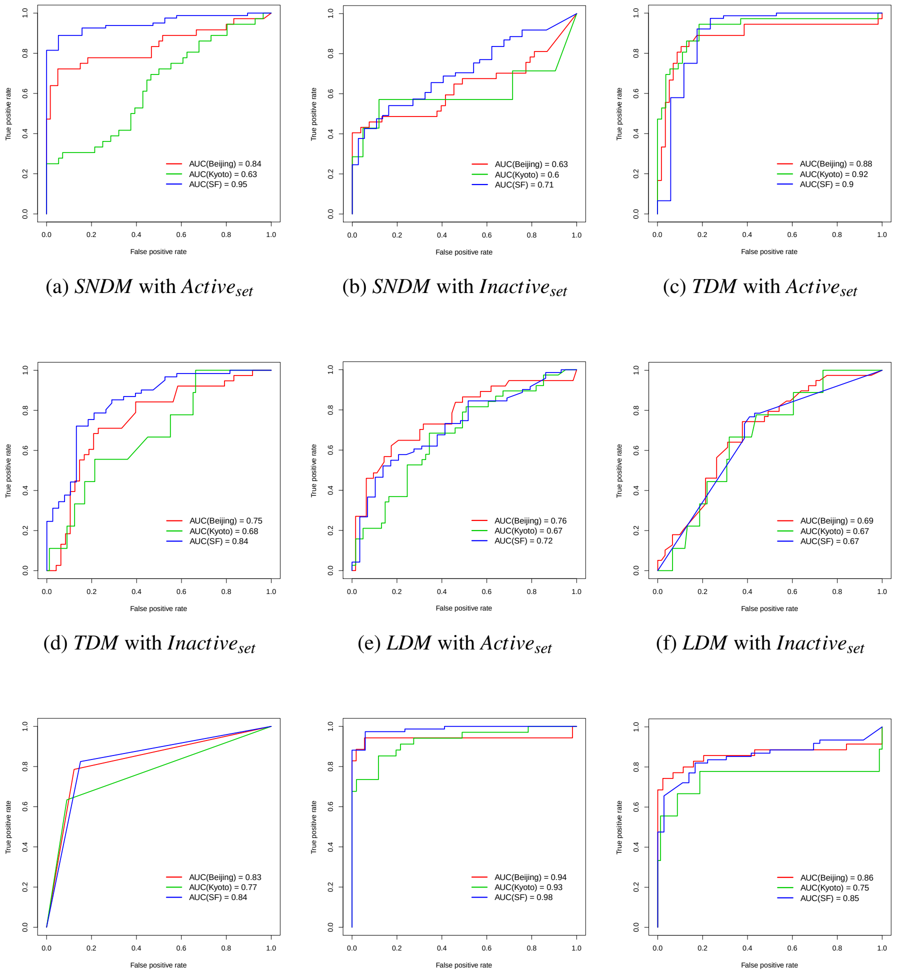
| **Datasets** | **One feature** | | | **Two features** | | | **Three features** |
|---|---|---|---|---|---|---|---|
| | SNDM | TDM | LDM | SNDM & TDM | SNDM & LDM | TDM & LDM | SNDM & TDM & LDM |
| $Active_{set}(Beijing)$ | 0.854 | 0.876 | 0.674 | 0.876 | 0.854 | 0.865 | **0.899** |
| $Active_{set}(Kyoto)$ | 0.682 | 0.847 | 0.647 | 0.824 | 0.694 | 0.847 | **0.847** |
| $Active_{set}(S.F.)$ | 0.828 | 0.882 | 0.641 | **0.935** | 0.774 | 0.871 | 0.892 |
| $Inactive_{set}(Beijing)$ | 0.747 | 0.709 | 0.620 | 0.772 | 0.747 | 0.684 | **0.772** |
| $Inactive_{set}(Kyoto)$ | 0.744 | 0.798 | 0.764 | 0.775 | 0.794 | 0.865 | **0.888** |
| $Inactive_{set}(S.F.)$ | 0.759 | 0.773 | 0.557 | **0.784** | 0.567 | 0.660 | 0.701 |

city, Figures 4 and 5 present the labeling variances between any two of the three subjects according to the Jaccard index. Indeed, all of the subjects provided high-quality outcomes, and their results were similar. Next, all of the models were analyzed using the ground truth from subjects' votes—i.e., by majority.

**Experimental Results:** The overall results for the location-authority estimation are reported in Figure 6. The results show that the proposed framework generally performed well. However, the results varied across models. The details are as follows.

*The social-network-driven model (SNDM) fluctuated considerably.* By comparing the ROC curves shown in Figures 6a and 6b, the decline in user activity had a negative effect on the classification results. Since inactive users always have few friends, SNDM is not so good at dealing with these users. The classification results shown in Figure 3 further verified this observation. For users with few friends, i.e., lower indegree and outdegree on the social network (see Definition 1), SNDM has insufficient discriminative power. Moreover, we analyzed the situations in different cities in order to gain insight into this model. By investigating the lower AUC values for Kyoto, we found that, because Kyoto is a traditional city, relatively few online social relationships are established there, compared to modern cities like San Francisco. In this sense, the performance of this model depends on different kinds of social culture.

*The time-driven model (TDM) always performed well.* Because the users in $active_{set}$ uploaded many images, this model obtained considerably good results that closely approximated the ground truth, as shown in Figure 6c. However, for users with few images, the model's performance decreased sharply. Figure 6d shows the model's performance when classifying inactive users. With Kyoto, for example, the low performance ($AUC = 0.68$) was due to calculations on a sparse matrix $M$, owing to the inactivity of users who upload few images on Flickr. Of the bottom 100 users from $Inactive_{set}(Kyoto)$, each of them uploaded an average of 89 images. In this sense, for users with few images, i.e., high sparsity of the visiting matrix (see Definition 2), TDM has insufficient discriminative power.



(a) *SNDM* with $Active_{set}$ (b) *SNDM* with $Inactive_{set}$ (c) *TDM* with $Active_{set}$

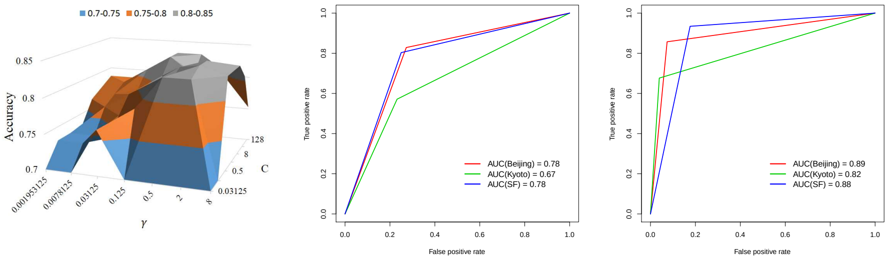
(d) *TDM* with $Inactive_{set}$ (e) *LDM* with $Active_{set}$ (f) *LDM* with $Inactive_{set}$

(g) General *SVM* (h) *DWE* with $Active_{set}$ (i) *DWE* with $Inactive_{set}$

**Fig. 6** Overall results for the location-authority-based user-classification task. By drawing the ROC curves, the five models' performance on both active (upper bound) and inactive (lower bound) datasets are presented.

*The location-driven model (LDM) returned the most stable results.* Figures 6e and 6f show the performance of the location-driven model. Although this model did not perform as well as the time-driven model, it did prove that location information is an important feature for stable results. That is, no matter what kind of raw data (e.g., how many friends, how many images, etc.) we collected for each user, LDM always performs stably. A more important characteristic is that combining it with other features improves significantly over using them alone, suggesting that the LDM adds important value to this classification task.

*The SVM model provides a solid baseline when integrating the above three models.* By using $Active_{set}$ and $Inactive_{set}$ as a whole, a three-fold cross validation was applied



(a) Tuning parameters for the *SVM* (b) Train: $Active_{set}$; Test: $Inactive_{set}$ (c) Train: $Inactive_{set}$; Test: $Active_{set}$

**Fig. 7** Detailed results for the SVM model: (a) the parameter-tuning process for obtaining the best classification model; (b) and (c) results from respectively using the $Active_{set}$ or $Inactive_{set}$ as the training set and the other for testing. Because the SVM model has only two outputs (high or low authority), there is only one threshold for the ROC curves.

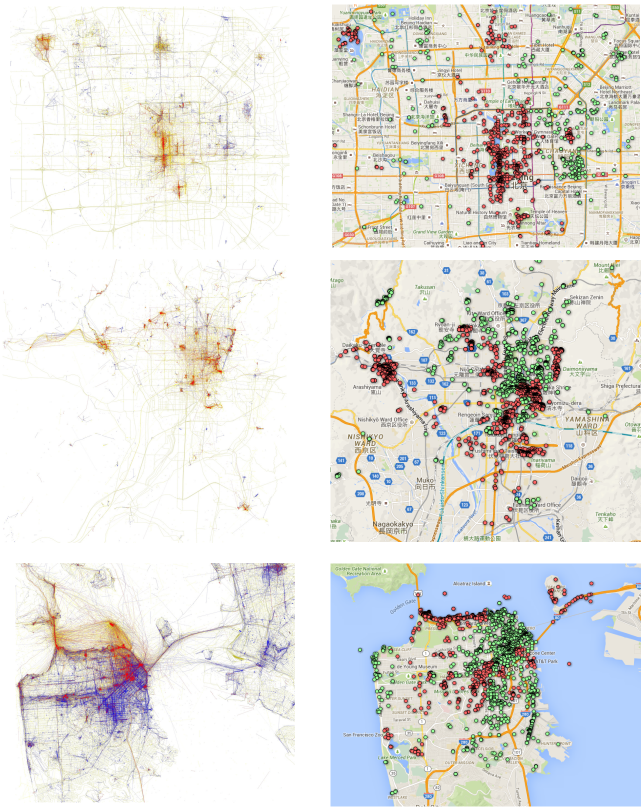
to train multiple SVM classification models. Figure 7a shows the grid-search method for tuning the parameters. The X-axis represents the RBF kernel's parameter $\gamma$, whereas the Y-axis denotes the penalty parameter $C$ in the SVMs. After choosing the best model from the parameter-tuning results, Figure 6g shows the general SVM model's superlative classification performance. Rather than randomly sampling training and testing data sets, Figures 7b and 7c show a cross-validation on $Active_{set}$ and $Inactive_{set}$. Because the model trained by $Inactive_{set}$ performed better than the model trained by $Active_{set}$, it is worthwhile exploring and implementing sophisticated features, rather than treating all of the features equally.

Table 7 further confirms that more features do not necessarily lead to better results. With San Francisco, for example, the SVM model obtained the best results when only two features were used. The relatively low performance from combining more features is due to the probability of existing disagreements between different models. This result has further verified our observation that SNDM, TDM, and LDM are good at dealing with different available raw data. Therefore, the DWE model is one of the best solutions to this problem.

*The DWE model performed best.* This model is an advanced version of the SVM model. The DWE model intelligently selects the most discriminative features for each user. By considering the differences in the available information for each user, dynamic weights were assigned to each feature. As a result, those models (SNDM, TDM, or LDM) that have sufficient discriminative power are automatically selected by DWE to classify each user. Ultimately, this model obtained the best results for both the $Active_{set}$ (Figure 6h) and $Inactive_{set}$ (Figure 6i).

## 4.4 Visualizing the Classification Results

Because our location-driven model, introduced in Section 3.3, was devised in accordance with the statistical observations of [10], we visualized our results to show that our model could generate similar distributions for the three cities Beijing, Kyoto, and San Francisco. Note that they used datasets that differ from ours, and neither our results nor theirs can be regarded as correct.



(a) Results in [10]

(b) Our results

**Fig. 8** Visualization of different user groups' mobility behavior in Beijing, Kyoto, and San Francisco. Temporal information is analyzed in (a), whereas location information is utilized in (b).

**Table 8** Representative users for the two latent groups.

| **Groups** | **Votes for the top ten representative members in the target group** | | |
|---|---|---|---|
| Tourist-like group | Beijing: 22/30; | Kyoto: 23/30; | San Francisco.: 26/30. |
| Local-like group | Beijing: 22/30; | Kyoto: 25/30; | San Francisco: 21/30. |

Although temporal information is an essential feature of such classifications, our results confirm that using only latitude–longitude coordinates can also perform well. Figure 8 visualizes the results. The red icons represent geo-tagged images taken by Flickr users who are very likely to be unfamiliar with the city, whereas the green icons represent images probably taken by those who are familiar with the city. Through such a classification, we can easily find popular sightseeing locations (i.e., the red areas) in each city. Thus, the location-driven model not only improves the time-driven model, it also models the relationships between people and space by introducing the inherent *authority* attribute.

To explain the latent variable *group*, Table 8 shows the number of votes obtained by the representative members from each group. The more votes a user has (with a maximum of three votes from the subjects), the more likely the user belongs to the corresponding group. By analyzing the votes in Table 8 and the distributions in Figure 8, we found that the representative members accurately reveal the underlying meaning of each group, i.e., a tourist-like group and a local-like group.

## 5 Application Scenario: Discovering Obscure POIs

In this section, we describe a real application for discovering relatively unknown POIs with the proposed user-profiling models. Unlike conventional studies that discover famous or popular POIs using a voting approach [2–5], because of the lack of related raw data on the Internet, our application discovers POIs that are not well known by analyzing information asymmetries among different user groups. This application well demonstrates the usefulness and novelty of the proposed models. In addition, because of the limited face-to-face access among social-media users scattered around the globe, we could not obtain a complete ground truth for the above evaluation in Section 4. Therefore, we conducted further indirect quantitative tests to verify whether the proposed user-profiling methods would help our application to discover obscure locations in Kyoto.

As mentioned in [1], obscure POIs are potential sightseeing resources. By dividing sightseeing locations into four quadrants on the basis of their “popularity” and “sightseeing quality” [30], this application aims to discover obscure POIs—i.e., those located in the quadrant with high sightseeing quality and low popularity. Obscure sightseeing locations are pertinent to in-depth travel, not only for enjoying the beautiful scenery but also for experiencing local culture, especially with repeat tourists who have already visited the more popular locations.

In accordance with these two dimensions, “popularity” and “sightseeing quality,” two main functions are implemented in our application: discovering obscure locations and determining whether they are worth visiting. In its current form, the application collects UGC from Flickr and supports user queries consisting of a target city $c$ and scenery object $o$ (e.g., “Kyoto, Maples”). The details for this application are as follows.

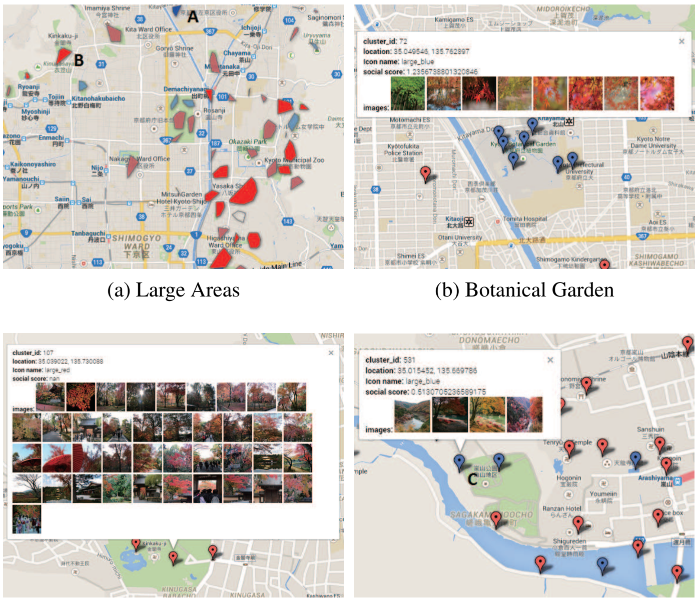
### 5.1 Discovering Obscure Locations

Our application first collects geo-tagged images and user profiles from Flickr using its abundant APIs [31]. By extracting friendship information, information regarding the time the images were captured, and geographic information, the respective data structures $N$ (Definition 1), $M$ (Definition 2), and $R$ (Definition 3) are established. Then, after applying all three user-profiling methods, the obscurity level of each sightseeing spot $s \in S$ (Definition 4) is calculated as follows by comparing the frequency of visits from different groups:

$$Obscurity(s) = \left\|\mathbf{vf}^{s}_{g=familiar}\right\| - \left\|\mathbf{vf}^{s}_{g=unfamiliar}\right\|. \tag{10}$$

The entries in vector $\mathbf{vf}^{s}$ represent the percentages of either the familiar fuzzy group or the unfamiliar fuzzy group who have visited spot $s$ each year. Basically, this equation means that if spot $s$ is visited more by unfamiliar users (e.g., foreign tourists), it is considered popular rather than obscure.

**Definition 4** A tree-based hierarchical graph $S$ is a collection of location clusters with a geographic hierarchical structure. In contrast to the GMM used in Definition 3, when a user changes the zoom level of the map, the hierarchical structure of $S$ can help our application



cluster_id: 72
location: 35.049546, 135.762897
Icon name: large_blue
social score: 1.2356738801320846
images:

(a) Large Areas (b) Botanical Garden

cluster_id: 107
location: 35.039022, 135.730088
Icon name: large_red
social score: nan
images:

cluster_id: 531
location: 35.015452, 135.669786
Icon name: large_blue
social score: 0.5130705236589175
images:

(c) Kinkaku-ji (d) Arashiyama

**Fig. 9** Interfaces for the proposed application, in which Google Maps' APIs are utilized. From deep blue to deep red, the color of each location represents how popular (or obscure) it is for sightseeing.

more efficiently display locations at different geographical granularities. Both density-based clustering algorithms (e.g., [32]) and agglomerative methods (e.g., hierarchical clustering) can be used for building a tree-based $S$. Because density-based clustering filters out a few sparsely distributed points that may be obscure, we utilized a centroid linkage-based hierarchical clustering algorithm to discover obscure spots.

## 5.2 Exploring Obscure Locations

Given the query "Kyoto, Maples," the application first automatically labels all related geotagged images of maple trees. A list of locations with maple trees in Kyoto is then displayed on a web-based map. By changing the zoom level of the map, users can explore popular and obscure locations at different geographical granularities.

Figure 9 shows the interfaces with multi-level representations of maple locations. Figure 9a shows the initial results. From deep blue (very obscure) to deep red (very popular), different colors indicate the obscurity levels of areas with maple trees. By zooming in on the map, Figures 9b and 9c show the Botanical Garden (i.e., an obscure spot labeled "A" in Figure 9a) and the Kinkaku-ji (i.e., a very popular spot, labeled "B" in Figure 9a), respectively. By clicking any tag on the map, users can view the corresponding images taken there.

In Figure 9d, some small obscure spots exist, even in a famous sightseeing area such as Arashiyama. The obscure maples in Arashiyama are located around parking lots and on mountains that are almost inaccessible to most foreign tourists. In this sense, we can conclude that more detailed user profiling is needed in order to discover fine-grained POIs.

**Table 9** Details of the cherry blossom dataset and the maple tree dataset.

| | **Cherry blossom dataset** | **Maple tree dataset** |
|---|---|---|
| # of geo-taged images | 4,576 | 6,950 |
| # of Flickr users | 319 | 177 |
| # of clustered spots in *S* (see Definition 4) | 368 | 582 |

Finally, the application evaluates whether visitors are satisfied with these obscure recommendations. We previously proposed two methods for judging scenery: a social-appreciation-based ranking and photographer-attention-based ranking. We detailed these methods in [33]. Because this paper mainly focuses on authority-based user profiling, the ability to detect obscure locations rather than recommend them is evaluated in the following section.

## 5.3 Evaluation of Obscure-location Discovery

In addition to the case studies described above, we further conduct the quantitative evaluation to verify the usefulness of our application. It also can be regarded as the indirect test of the proposed user-profiling models. Therefore, an indirect test is conducted to verify whether the user-profiling models would help our application to discover obscure locations in Kyoto.

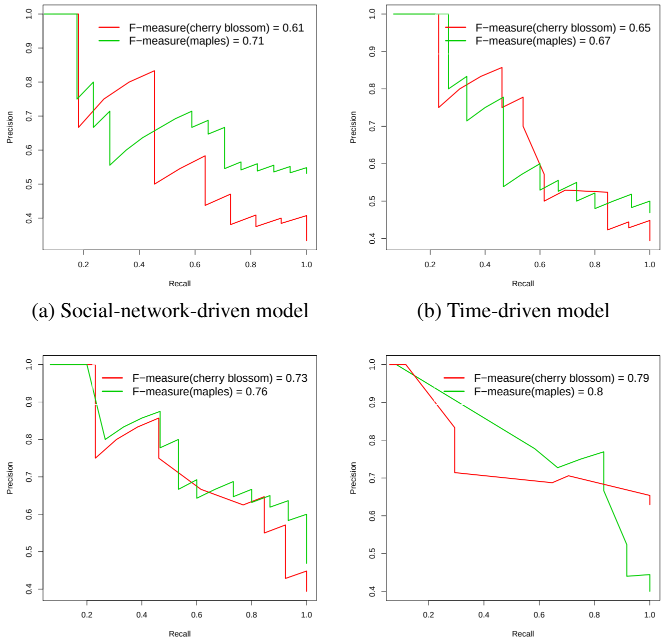
**User Queries:** In accordance with the input for our application, we selected two user queries: “Kyoto, Cherry Blossoms” and “Kyoto, Maples.” First, 929,403 geo-tagged images were crawled from Flickr. By using Flickr’s keyword-based search and our labeling component, 4,576 geo-tagged images of cherry blossoms in Kyoto and 6,950 geo-tagged images of maple trees in Kyoto were labeled.

Table 9 summarizes the two target datasets. In total, 4,576 cherry blossom images were uploaded by 319 Flickr users, while the other 177 users contributed 6,950 images of maple trees. By generating the leaves of the tree-based graph *S* (see Definition 4), we obtained 368 cherry blossom locations and 582 maple locations in Kyoto.

**Ground Truth:** Before analyzing the effect of our application, we first invited three residents who had lived in Kyoto for more than 20 years to label the discovered locations using a five-point scale ranging from “0” (not famous) to “4” (very famous). To reduce their workload, we selected 33 cherry-blossom candidate locations and 32 maple candidate locations, for which 17 and 18 obscure locations were returned, respectively, by regarding the average scores of the three respondents with a relevance score of no more than 2.0.

**Experimental Results:** Because the location-driven model cannot be applied alone, Figure 10 shows the precision-recall curves from Eq. 10’s experimental results obtained by using the other four methods. The red curve shows the experimental results from the query “Kyoto, Cherry Blossom.” The green curve displays the corresponding results from the query “Kyoto, Maples.” Precision represents the fraction of discovered spots that are relevant, whereas recall represents the fraction of relevant spots that were discovered. Given these results, we can make the following two conclusions.

*Location authority is useful for discovering obscure locations.* We noticed that conventional authority-based methods such as “rank-by-visits” and “rank-by-users” cannot efficiently detect obscure locations. For example, the obscure botanical garden in Figure 9b may be wrongly identified as a popular location because of the high accumulation of images taken by locals or professional photographers. Without introducing users’ authority information, the obscure spots located in popular areas cannot be found (see the blue tags in



(a) Social-network-driven model

(b) Time-driven model

(c) Ensemble 1: SVM

(d) Ensemble 2: DWE

**Fig. 10** Precision-recall curves for discovering obscure cherry blossoms and maple trees in Kyoto. Each query's F-measure is also computed and shown in this figure.

Figure 9d). Having obtained a high F-measure (0.72 on average), our methods effectively reveal that the "obscurity" is caused by the information asymmetry among different user groups. Thus, our authority-based solution is capable of identifying such groups.

*The temporal information used in both our time-driven model and Fischer's method [10] is insufficient.* Figure 10b shows that the time-driven model (Eq. 2) did not outperform the other methods in a real scenario. The reason for this is that only approximately 20% of the Flickr users uploaded and shared several images on the basis of the statistics in Section 4.1. Insofar as they are compatible with differently available information, the ensembles yielded better results, especially the dynamic weighted ensemble shown in Figure 10d. Although the time-driven model performed better in the experiments in Section 4.3, the other two complementary models are needed for robustness.

## 6 Conclusions

In this paper, we proposed a new concept, *location authority*, for mining information asymmetries among different online user groups. By respectively considering a user's social, temporal, and spatial information, we devised three models for estimating users' location authority: a social-network-driven model, time-driven model, and location-driven model. We then introduced two ensembles of these models, which demonstrated robust performance across different classification tasks. Furthermore, by utilizing our profiling solution, a real application was implemented in order to discover obscure sightseeing locations in Kyoto.

In the experiments, we manually collected the ground truth and compared the performance of each proposed method. By investigating the obscure locations detected with our application, we also conducted an indirect evaluation, and visualized the experimental results in order to show the broader applications of our methods. Our work is intended to inspire more interest in information asymmetry analysis, user profiling, and geo-social recommendations.

## References

1. Y. Zheng, Z. Zha, and T. Chua, "Research and applications on georeferenced multimedia: a survey," *Multimedia Tools Appl.*, vol. 51, no. 1, pp. 77–98, 2011.
2. W.-C. Chen, A. Battestini, N. Gelfand, and V. Setlur, "Visual summaries of popular landmarks from community photo collections," in *ACM Multimedia*, 2009, pp. 789–792.
3. Q. Hao, R. Cai, C. Wang, R. Xiao, J. Yang, Y. Pang, and L. Zhang, "Equip tourists with knowledge mined from travelogues," in *WWW*, 2010, pp. 401–410.
4. R. Ji, X. Xie, H. Yao, and W.-Y. Ma, "Mining city landmarks from blogs by graph modeling," in *ACM Multimedia*, 2009, pp. 105–114.
5. J. Liu, Z. Huang, L. Chen, H. T. Shen, and Z. Yan, "Discovering areas of interest with geo-tagged images and check-ins," in *ACM Multimedia*, 2012, pp. 589–598.
6. H.-W. Chang, D. Lee, M. Eltaher, and J. Lee, "@ phillies tweeting from philly? predicting twitter user locations with spatial word usage," in *ASONAM*, 2012, pp. 111–118.
7. M. D. Lieberman and J. Lin, "You are where you edit: Locating wikipedia contributors through edit histories." in *ICWSM*, 2009, pp. 106–113.
8. A. Popescu and G. Grefenstette, "Mining user home location and gender from flickr tags." in *ICWSM*, 2010, pp. 307–310.
9. L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: improving geographical prediction with social and spatial proximity," in *WWW*, 2010, pp. 61–70.
10. https://www.flickr.com/photos/walkingsf/sets/72157624209158632/.
11. J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," in *EMNLP*, 2010, pp. 1277–1287.
12. H. Gao, J. Tang, and H. Liu, "Exploring social-historical ties on location-based social networks," in *ICWSM*, 2012, pp. 114–121.
13. E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *SIGKDD*, 2011, pp. 1082–1090.
14. D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *SMUC*, 2010, pp. 37–44.
15. M. Pennacchiotti and A.-M. Popescu, "Democrats, republicans and starbucks afficionados: user classification in twitter," in *SIGKDD*, 2011, pp. 430–438.
16. J. Luo, D. Joshi, J. Yu, and A. Gallagher, "Geotagging in multimedia and computer vision - a survey," *Multimedia Tools Appl.*, vol. 51, no. 1, pp. 187–211, 2011.
17. Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *WWW*, 2009, pp. 791–800.
18. Y. Zheng and X. Xie, "Learning travel recommendations from user-generated gps traces," *ACM TIST*, vol. 2, no. 1, p. 2, 2011.
19. K. Hasegawa, Q. Ma, and M. Yoshikawa, "Trip tweets search by considering spatio-temporal continuity of user behavior," in *DEXA (2)*, 2012, pp. 141–155.
20. D. Xu, P. Cui, W. Zhu, and S. Yang, "Find you from your friends: Graph-based residence location prediction for users in social media," in *ICME*, 2014, pp. 1–6.
21. Z. Gyöngyi, H. Garcia-Molina, and J. O. Pedersen, "Combating web spam with trustrank," in *VLDB*, 2004, pp. 576–587.
22. C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.
23. C. wei Hsu, C. chung Chang, and C. jen Lin, "A practical guide to support vector classification," 2010.
24. L. Rokach, *Pattern Classification Using Ensemble Methods*. River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2010.
25. L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
26. J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class adaboost," *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

27. D. Jiménez, “Dynamically weighted ensemble neural networks for classification,” in *IJCNN*, 1998, pp. 753–756.
28. L. Pretto, “A theoretical analysis of googles pagerank,” in *SPIRE*, 2002, pp. 131–144.
29. T. L. Griffiths and M. Steyvers, “Finding scientific topics,” *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.
30. C. Zhuang, Q. Ma, X. Liang, and M. Yoshikawa, “Anaba: An obscure sightseeing spots discovering system,” in *ICME*, 2014, pp. 1–6.
31. https://www.flickr.com/services/api/.
32. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *SIGKDD*, 1996, pp. 226–231.
33. C. Zhuang, Q. Ma, X. Liang, and M. Yoshikawa, “Discovering obscure sightseeing spots by analysis of geo-tagged social images,” in *ASONAM*, 2015, pp. 1–6.