Where you Instagram? Associating Your Instagram Photos with Points of Interest

Xutao Li[+] Tuan-Anh Nguyen Pham[+] Gao Cong[+]
Quan Yuan[+] Xiao-Li Li[#] Shonali Krishnaswamy[#]
[+]School of Computer Engineering, Nanyang Technological University, Singapore.
{lixutao@, pham0070@e., gaocong@, qyuan1@e.}ntu.edu.sg
[#]Institute for Infocomm Research(I2R), A*STAR, Singapore.
{xlli, spkrishna}@i2r.a-star.edu.sg

ABSTRACT

Instagram, an online photo-sharing platform, has gained increasing popularity. It allows users to take photos, apply digital filters and share them with friends instantaneously by using mobile devices. Instagram provides users with the functionality to associate their photos with points of interest, and it thus becomes feasible to study the association between points of interest and Instagram photos. However, no previous work studies the association. In this paper, we propose to study the problem of mapping Instagram photos to points of interest. To understand the problem, we analyze Instagram datasets, and report our findings, which also characterize the challenges of the problem. To address the challenges, we propose to model the mapping problem as a ranking problem, and develop a method to learn a ranking function by exploiting the textual, visual and user information of photos. To maximize the prediction effectiveness for textual and visual information, and incorporate the users' visiting preferences, we propose three subobjectives for learning the parameters of the proposed ranking function. Experimental results on two sets of Instagram data show that the proposed method substantially outperforms existing methods that are adapted to handle the problem.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Filtering

Keywords

Photo Mapping; Ranking; Point of Interest

1. INTRODUCTION

Instagram, a mobile based photo-sharing system, allows users to share their photos with friends instantaneously on various social networking platforms, e.g., Facebook, Twitter, etc. It provides users with many digital filters for creating magic visual effects on their photos, and users can express their sentiments/opinions and give comments on uploaded photos. Launched in October 2010, Instagram has gained popularity rapidly over the past several years. As of December 2014, it had over 300 million active users who shared more than 20 billion photos.[1] The Instagram community is still growing at a rapid speed and on average more than 40 million photos are uploaded per day.

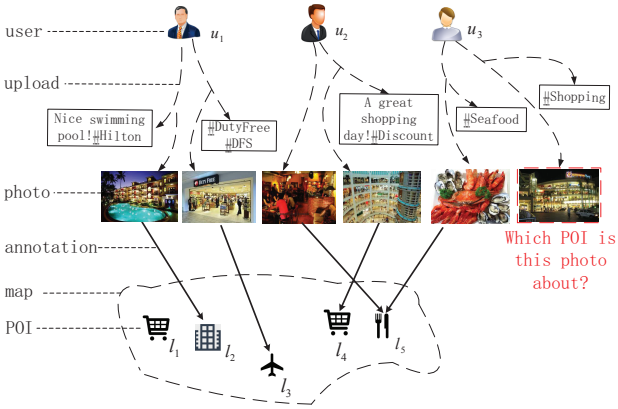


Figure 1: A graph representation of the interactions in Instagram.

One interesting functionality of Instragram is that it allows users to associate photos with Points of Interest (POIs), which offers an exciting opportunity to study the association between photos and POIs. In Figure 1, we give an example to show how users, photos and POIs interplay with each other. In Instagram a user can upload photos with textual descriptions and hashtags. For example, user u_1 submitted a photo with the description "Nice swimming pool!" and meanwhile the hashtag "#Hilton" was used. Users can also associate the uploaded photos with POIs, e.g., user u_1 annotated the two pictures he uploaded to the POIs ℓ_2 and ℓ_3, respectively.

Knowing the association between photos and POIs has important applications. First, it provides both visual and textual annotations about POIs for users to explore. For example, when we are interested in visiting a POI, we can browse the photos associated with it to see the POI and it-

[1]http://www.statista.com/statistics/253577/number-of-monthly-active-instagram-users/

CIKM'15, October 19 - 23, 2015, Melbourne, VIC, Australia

DOI: http://dx.doi.org/10.1145/2806416.2806463 .

s environment; meanwhile, we can read the corresponding textual information to find interesting aspects or social activities related to the POI. Second, the association is a fundamental data preprocessing task, which is a prerequisite for other mining tasks. Third, associating photos and POIs will help to understand users' preferences over POIs. Recently, Silva et al. have studied the Instagram data [22], compared Instagram photos with the check-ins on Foursquare [23], and shown that photos can be mimicked as check-ins at POIs.

When users upload photos, they need manually associate photos with POIs, which can be tedious. In Instagram, only less than 30% pictures are annotated to POIs by users. Thus, studying how to automatically map Instagram photos to POIs is an important problem. However, to the best of our knowledge, no existing work has considered the problem.

In this paper, we propose to study the problem of mapping the Instagram photos to POIs. Specifically, given the photos that have been associated with POIs as training set, we learn a mapping function to annotate a photo that is not associated with POI. The problem is challenging because of two reasons. First, as we will see in Section 3.2, we find that the number of photos associated with each POI follows a power law distribution. That is, a great number of POIs have very few photos and thus it is difficult to train an effective mapping model for them due to the data scarcity problem. Second, we have not only visual and textual information, but also user visiting information to explore. The visiting preferences of users over POIs, as an important type of information, should be exploited, which is also a way to overcome the data scarcity problem. As we will see in Section 3.2, however, we find that users tend to take photos at POIs that are new to them, and this makes it difficult to exploit users' visiting preferences.

Note that our problem differs from previous studies on mapping Flickr photos to the earth. Previous work on Flickr photos can be classified into two types. One type is associating photos to some landmarks [5, 14, 3], and the other is dividing the world map into grids and then predicting which grid a photo belongs to [20, 18, 11, 16]. For the first type, the landmarks usually refer to popular places and thus there are enough photos in each landmark to train a model. In our case, as aforementioned we lack photos for the majority of POIs due to the power law distribution. Furthermore, in these approaches [5, 14], the number of landmarks considered is limited, and it is unclear how to effectively scale these approaches coping with many POIs when the training samples are scarce. For the second type, a grid region usually contains multiple POIs and mapping photos to a grid level is insufficient to annotate POIs. Another major difference is these previous studies do not exploit the visiting preferences of users to POIs. However, Instagram data reveals users' visiting preferences over POIs [23], and thus it is important to exploit user information in our problem.

To solve the photo mapping problem, we first explore the characteristics of Instagram data. We make three interesting observations: (1) most of POIs have fewer than 10 photos; (2) users in general associate fewer than 50 photos with POIs in one year and the number of users who associate more than 50 photos decreases in a power law fashion; (3) users tend to take photos at POIs that are new to them. These findings are reported for the first time, and also characterize the challenges of our problem (to be analyzed in Section 3.2). Second, we model the mapping problem as a ranking problem, and define a function to score POIs in terms of the textual context, visual content and user information for a given photo. In particular, we propose a new multi-task model to combine the textual, visual and user information for learning the scoring function. In the proposed model, we maximize prediction abilities of the combined (textual, visual and user) information as the main objective, and maximize prediction effectiveness of textual and visual information as two subobjectives. As a result, the most informative features from textual context and visual content can be effectively combined into the function. Moreover, we effectively incorporate the visiting preferences of users into the scoring function by designing another weighted matrix factorization subobjective. The contributions of this paper can be summarized as follows.

- We define the new problem of associating Instagram photos with POIs. We make interesting observations on Instagram and report new findings on Instagram for the first time. To the best of our knowledge, we are the first to propose and study the problem.

- We propose a method, Ranking with Textual, Visual and User information for Photos to POIs, called Rank-TVU_P2P, which is able to combine not only the textual and visual information but also users' visiting preferences for mapping Instagram photos to POIs. In the proposed method, we develop a subtask to maximize the prediction effectiveness for each type of information (textual context, visual content and user). In particular, a weighted matrix factorization subtask is developed for modeling the visiting preferences of users. As a result, the most effective information is combined into our scoring function for POIs.

- Extensive experiments on two sets of Instagram data, namely, New York City and Singapore, are conducted, and the results show that our proposed method outperforms baseline methods significantly. Moreover, the experimental results demonstrate the usefulness of each type of information, and the importance of our developed subobjectives (subtasks).

The rest of the paper is organized as follows. In Section 2, we briefly review related work. In Section 3, we explore the Instagram data and present some interesting patterns we find. Our proposed model is introduced in Section 4 and experimental results are presented in Section 5. Finally, we conclude the paper with some future research directions in Section 6.

2. RELATED WORK

2.1 Geolocating Images

Previous work on geolocating images can be categorized into two types. One type is dividing the concerned region into grids and predicting the grid where an image resides, and the other type is identifying from a set of candidates the landmark that an image refers to.

Grid Prediction. Hays and Efros [6] propose a data driven approach to calibrating Flickr images to grids on the earth. In the approach, they first filter Flickr images by some specific geo-related and some obviously geo-unrelated

tags, and then place the images onto $200km \times 200km$ grids by employing K-nearest neighbors algorithm with some predefined visual features. Instead of leveraging visual features, Serdyukov et al. [20] adopt textual features, i.e., user-generated tags, to map the Flickr photos to grids. They build a Language Model (LM) for each grid in terms of its associated tags, and place the photos according to the probability calculated by the LMs. Liu et al. [16] propose to build LM for a user in terms of the tags he/she used, and then combine it with grid-based LM for mapping Flickr images to grids. In this approach, only textual features are considered.

Our problem differs from these previous studies [6, 20, 16] mainly in two aspects. First, we aim to associate Instagram photos with POIs, instead of grids. A grid usually contains multiple POIs and mapping photos to grid level is insufficient to annotate POIs. Nevertheless, we adapt the approach in the work [16] for our problem, and compare with it in our experiments. Second, previous studies do not consider the visiting preferences of users. However, the photos uploaded in Instagram resemble check-ins [23], which are very useful to characterize users' preferences over POIs and should not be overlooked for our problem.

Landmark Identification. Crandall et al. [5] combine both visual (Scale Invariant Feature Transform–SIFT features [17]) and textual (tag features) information to place Flickr photos onto landmarks, which is a representative work on landmark identification. In their work, they focus on identifying which of 10 landmarks in a city an image belongs to. Their solution is considering the problem as a multi-class classification problem. In particular, they build Support Vector Machine (SVM) and Naive Bayesian (NB) based on visual and textual features, respectively, and then combine the classification results of them to produce final prediction. Other example studies on landmark identification include [14], [3], [25] and [1].

Our work differs from these studies in two aspects. First, the landmarks often refer to popular places and thus there exit sufficient photos to train a classification model. However, in our problem, we lack training photos because most of POIs have only a few photos. Moreover, the number of landmark candidates considered in previous studies is very small , e.g., 10 landmarks [5], while the number of POIs is much larger, e.g., more than one thousand. It is not very effective to train a classification model for our problem where the training samples are scarce, as we will see in the experiments. Second, these studies do not consider the users' preferences over POIs.

2.2 Geolocating Tweets or Users

As a vast amount of tweets and online information are generated by users, several studies have been conducted for predicting locations of tweets or users [2, 4, 12, 24]. For example, based on user-supplied home address information and social friendship information in Facebook, Backstrom et al. [2] propose to predict the locations of users by measuring and using their social and spatial proximity. Cheng et al. [4] develop a probabilistic approach to estimating the city-level location of a Twitter user, which purely relies on the content of his/her tweets. In their approach, a classification model is first built to identify the geoscope word in tweets for estimating users' locations, and then a lattice-based neighborhood smoothing model is proposed to refine the estimations. Recently, Li et al. [12] study a similar problem to identify the cities of residence for Twitter users. They propose a unified discriminative influence model based on the assumption that Twitter users tend to follow users living close to them. However, these approaches focus on a city-level granularity and cannot solve our problem because city-level accuracy is too coarse to associate photos with POIs. Moreover, users' visiting preferences over POIs and visual features of photos are not explored in the approaches. Yuan et al. [24] develop a spatio-temporal topic model based on Twitter data, which can also be used for estimating users' locations or tweets' locations given a time.

Table 1: Data statistics of NYC and SG.

Statistics	NYC	SG
# of photos	21,910,375	13,168,666
# of POIs	602,604	372,104
# of users	126,543	87,281
% of photos annotated to POIs	29	26
% of photos with textual info.	92	94

3. INSTAGRAM EXPLORATION

3.1 Data Description

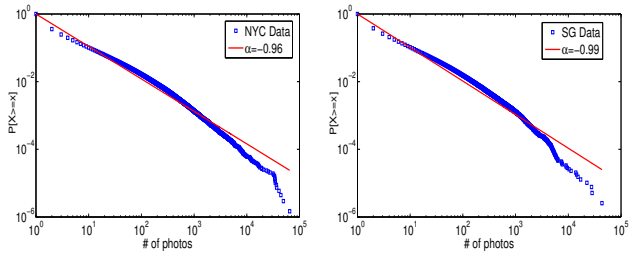
We crawled two sets of Instagram data from 13 Nov 2013 to 13 Nov 2014, which were from users in New York City (NYC) and Singapore (SG), respectively. When crawling the data, a user was considered from NYC or SG if more than a half of his/her Instagram photos were taken in the region of NYC or SG. We refer to the two datasets as NYC and SG, respectively. NYC comprises 602,604 POIs and 21,910,375 photos taken by 126,543 users, and SG comprises 372,104 POIs and 13,168,666 photos taken by 87,281 users. 92% and 94% of NYC and SG photos are accompanied with textual information, respectively. When users upload photos, they need manually associate photos with POIs. We find that 29% and 26% photos have been already annotated to POIs by users for both data, respectively. In other words, more than 70% of photos in Instagram are not mapped to POIs. The basic statistics of both data are summarized in Table 1.

3.2 Data Analysis

Observation 1: Number of photos per POI. We show in Figures 2(a) and 2(b) the empirical Complementary Cumulative Distribution Function (CCDF) of number of photos per POI, for NYC and SG data, respectively. We observe that both CCDFs follow the power law distribution, i.e., $P[X \geq x] = x^{\alpha}$, where X is a random variable for the number of photos and α is a coefficient. By fitting the empirical distribution, we obtain the coefficient $\alpha = -0.96$ and $\alpha = -0.99$ for NYC and SG, respectively.

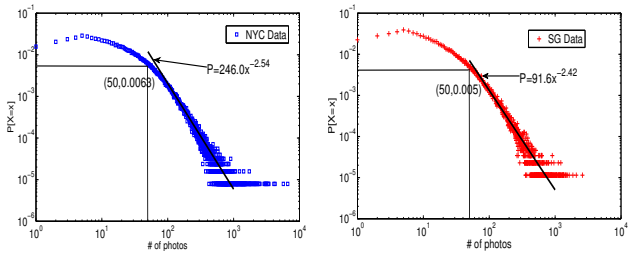
Given a power law distribution, we can compute $x \approx \lfloor \exp(\frac{\ln P}{\alpha}) \rfloor$ if we know the probability $P[X \geq x]$. For example, when $P[X \geq x] = 10\%$, we have $x \approx 11$ and $x \approx 10$ for NYC and SG, respectively.[2] This means only 10% of POIs have at least 11 or 10 photos for NYC and SG, respectively, and the other 90% of POIs have no more than 11 or 10 photos. Due to this issue, it would not be effective to train a

[2]This is in accordance with the empirical distribution we computed, where $P[X \geq 11] = 9.82\%$ and $P[X \geq 10] = 10.2\%$ for NYC and SG, respectively.



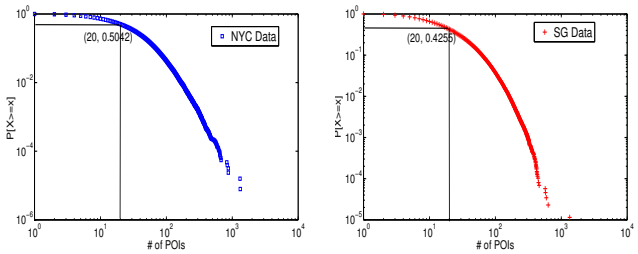
(a) NYC (b) SG

Figure 2: Distribution of number of photos per POI.



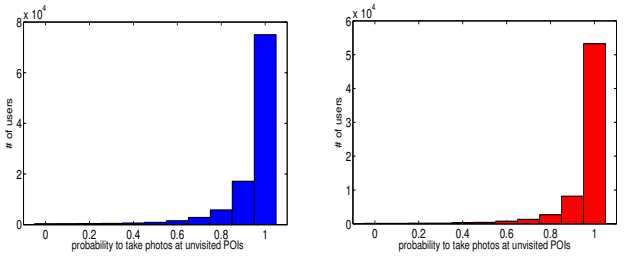
(a) NYC (b) SG

Figure 3: Distribution of number of photos associated with POIs per user.



(a) NYC (b) SG

Figure 4: Distribution of number of POIs per user.



(a) NYC (b) SG

Figure 5: Histogram of the probability to take photos at POIs that are new to users.

classifier for POIs as the previous methods [14, 5] for Landmark identification do (as we will see in the experiments), because most of POIs will lack training samples.

Observation 2: Number of photos per user. We show in Figures 3(a) and 3(b) the empirical Probability Density Function (PDF) of number of photos associated with POIs per user. We can see from both figures that the probability first slightly increases, peaks when the number of photos is around 5, and then decreases rapidly after 50, which is not a power law distribution.[3] However, the righthand part after the peak has a heavy tail, which can be modeled with a power law distribution, as shown in the figures. This observation indicates that Instagram users in general associate fewer than 50 photos with POIs in one year (recall that both datasets are for one year period), and the number of users who associate more than 50 photos decreases with a power law probability. This observation suggests that we have only a few annotated training samples for each user.

Observation 3: Behaviors of users. First, we study the number of POIs where a user took photos. We plot the empirical CCDF of number of POIs per user in Figures 4(a) and 4(b) for NYC and SG data, respectively. We observe that 50.42% of users in NYC and 42.55% of users in SG have visited and taken photos at more than 20 POIs. This indicates we cannot easily determine which POI a photo is associated with based on users' visiting information, because users take photos at many POIs.

Another interesting finding about users' behaviors is that they incline to take photos at POIs that are new to them. To observe this, we calculate the probability that a user takes photos at newly-visited POIs by the following three steps: (i) for each user, we sort his/her photos taken at (associated with) POIs by time; (ii) we employ the first five POIs as initial visited set for each user, and then iterate through his/her sorting list to count whether a photo is taken at a new POI or a visited POI; meanwhile, we update the visited POI set by including the POI for each checked photo; (iii) according to the counting results, we can compute the probability P_u for user u to take a photo at new POIs as:

$$P_u = \frac{n_u}{n_u + o_u} \tag{1}$$

where n_u counts the number of new POIs for user u in step (ii) and o_u counts the number of visited POIs. Finally, we depict a histogram with the probabilities of all the users. The histograms are shown in Figures 5(a) and 5(b) for NYC and SG, respectively.

We can see that a greater portion of users have high probability (say 0.8 to 1.0) to take photos at newly-visited POIs (or we call previously-unvisited POIs for a user). This can be explained by the intuition that new POIs are usually more attractive for users to take photos than the visited ones; for previously visited POIs, users rarely take photos and annotate them again within the period of our data collection (1 year). The finding further indicates the difficulty in exploiting users' visiting behaviors to associate photos with POIs.

4. PROBLEM DEFINITION AND PROPOSED METHOD

In this section, we first define the problem of associating Instagram photos with POIs, and then present our proposed approach to solving the problem.

4.1 Notations and Problem Definition

Notations. Let $\mathcal{U}$ be a set of users $\{u_1, u_2, \cdots, u_{|\mathcal{U}|}\}$, $\mathcal{P}$ denote a set of photos $\{p_1, p_2, \cdots, p_{|\mathcal{P}|}\}$ and $\mathcal{L}$ denote

[3]It resembles a Double Pareto Lognormal Distribution [21].

Table 2: A list of notations

symbols	meanings
$\mathcal{U}$	the set of users: $\{u_1, u_2, \cdots, u_{\|\mathcal{U}\|}\}$
$\mathcal{P}$	the set of photos: $\{p_1, p_2, \cdots, p_{\|\mathcal{P}\|}\}$
$\mathcal{L}$	the set of POIs: $\{\ell_1, \ell_2, \cdots, \ell_{\|\mathcal{L}\|}\}$
$\mathbf{x}_p$	textual feature vector of photo p
$\mathbf{y}_p$	visual feature vector of photo p
$\mathcal{D}$	training set $\mathcal{D} \subset \mathcal{P}$, where $p \in \mathcal{D}$ has $p.u$ and $p.\ell$
$\mathbf{w}_\ell$	textual factor of POI ℓ interacting with photos
$\mathbf{v}_\ell$	visual factor of POI ℓ interacting with photos
$\mathbf{u}_u$	latent factor of user u interacting with POIs
$\mathbf{l}_l$	latent factor of POI ℓ interacting with users
$h(z)$	the hinge-loss function: $max(0, 1-z)$

a set of POIs $\{\ell_1, \ell_2, \cdots, \ell_{|\mathcal{L}|}\}$ in an Instagram database. As photos are associated with textual context, we employ the bag-of-word model to represent each photo p as an m-dimensional vector $\mathbf{x}_p$, where x_{pi} denotes the frequency of word i for photo p. We note that $\mathbf{x}_p$ is a zero vector if a photo does not have textual description. However, this does not take place usually because more than 90% photos have the textual information as aforementioned in Table 1. For visual content, we extract visual words to construct features [5]. Specifically, for each photo, we first use the standard computer vision technique Scale Invariant Feature Transform (SIFT) [17][4] to produce a set of keypoints, where each keypoint is represented as a 128-dimensional vector. Subsequently, the keypoints from all the photos are clustered into n-clusters to create a "visual vocabulary", where each cluster is considered as a "visual word". Finally, we assign each keypoint in a photo to its closest cluster, and then represent each photo as a n-dimensional vector which indicates how many times each "visual word" appears in the photo. That is, for each photo p, we form a n-dimensional vector $\mathbf{y}_p$, where y_{pi} denotes the frequency of "visual word" i appearing in photo p. For clarity, we summarize the notations used in this paper in Table 2.

Associating photos with POIs: *Based on the set of photos that are associated with POIs by Instagram users, we aim at learning a function for mapping a photo to a POI. Formally, given a training set $\mathcal{D} \subset \mathcal{P}$, where each $p \in \mathcal{D}$ has two attributes $p.u$ and $p.\ell$: $p.u$ denoting the user who uploads photo p and $p.\ell$ denoting the POI that photo p is associated with, we learn a mapping function $f : \{(u, p, \mathbf{x}_p, \mathbf{y}_p)\} \rightarrow \mathcal{L}$ which is used to predict the POI for a photo that is not associated with a POI.*

According to **Observation** 1 in Section 3.2, the number of training samples for most of POIs is small. Hence, building a classifier for POIs in $\mathcal{L}$ based on textual and visual features may not yield good accuracy for mapping. To learn the mapping function we propose a new approach that effectively exploits not only the visual and textual information of photos, but also the information of the users who upload them. Next, we present our proposed method.

4.2 Proposed Method

We approach the mapping problem by treat it as a ranking problem. We develop an approach to learning a function from the training set $\mathcal{D}$ to score each POI for a given photo to be associated, and then rank the POIs based on their scores. The textual context, visual content, and user information of a photo are taken into account in learning the ranking function. Since there are multiple types of information, we design a subtask for each type to maximize its predicting effectiveness. As a result, the most effective information of each type can be combined into the scoring function.

Next, we present how we embody this idea in details. We first embody our method by only considering textual context and visual content, and then introduce how our method exploits the user information.

4.2.1 Combining Textual Context and Visual Content

In this subsection, we present an approach to combining textual contexts and visual contents of photos to learn a scoring function for mapping a photo to a POI. We first introduce two factors $\mathbf{w}_\ell$ and $\mathbf{v}_\ell$ of m-dimension and n-dimension for each POI ℓ, and then use them to model the interactions with textual and visual features of photos, respectively. The scoring function is defined as follows:

$$s_{p\ell} = \mathbf{w}_\ell \cdot \mathbf{x}_p + \mathbf{v}_\ell \cdot \mathbf{y}_p \tag{2}$$

where $s_{p\ell}$ denotes the interaction score between photo p and POI ℓ, and operator $\cdot$ represents the inner product. Clearly, the score $s_{p\ell}$ is computed by a sum of textual interaction score $\mathbf{w}_\ell \cdot \mathbf{x}_p$ and visual interaction score $\mathbf{v}_\ell \cdot \mathbf{y}_p$ between POI ℓ and photo p.

Next, our objective is to learn $\{\mathbf{w}_\ell\}_{\ell \in \mathcal{L}}$ and $\{\mathbf{v}_\ell\}_{\ell \in \mathcal{L}}$ such that for each photo p its correct POI is expected to be ranked higher than the other POIs. In particular, we extend *a pairwise classification* approach that is used in RankSVM [9] for this objective. That is, given each training sample $p \in \mathcal{D}$, we compare pairwisely the scores of POIs $p.\ell$ and ℓ' ($\ell' \in \mathcal{L}$ and $\ell' \neq p.\ell$) to check whether they are ranked correctly; if $p.\ell$ is ranked higher than ℓ', there is no loss; otherwise a loss is produced. Specifically, we have the following objective function to minimize:

$$\mathcal{O}_{main} = \sum_{\substack{p \in \mathcal{D}, \\ \ell = p.\ell}} \sum_{\substack{\ell' \in \mathcal{L}, \\ \ell' \neq \ell}} h(s_{p\ell} - s_{p\ell'}) \tag{3}$$

where $h(z) = max(0, 1-z)$ is the hinge-loss for converting the score difference between two POIs into a penalty. We note that the hinge-loss is nonzero only if $z < 1$.

In Eq. (3) the two types of information, i.e., the textual context and the visual content, are combined directly and the objective function might not exploit the most informative features from each type of information. For overcoming this issue, we design two subtasks for assistance, namely, to rank the correct POI higher than the other POIs when only a single type of information is presented. By following the idea of Eq. (3), we write another subobjective function to minimize for each subtask:

$$\mathcal{O}_1 = \sum_{\substack{p \in \mathcal{D}, \\ \ell = p.\ell}} \sum_{\substack{\ell' \in \mathcal{L}, \\ \ell' \neq \ell}} h(\mathbf{w}_\ell \cdot \mathbf{x}_p - \mathbf{w}_{\ell'} \cdot \mathbf{x}_p) \tag{4}$$

and

$$\mathcal{O}_2 = \sum_{\substack{p \in \mathcal{D}, \\ \ell = p.\ell}} \sum_{\substack{\ell' \in \mathcal{L}, \\ \ell' \neq \ell}} h(\mathbf{v}_\ell \cdot \mathbf{y}_p - \mathbf{v}_{\ell'} \cdot \mathbf{y}_p). \tag{5}$$

When minimizing $\mathcal{O}_1$, the factor $\mathbf{w}_\ell$ will encode the most discriminative textual features for POI ℓ, as $\mathcal{O}_1$ aims at ranking POIs correctly by only utilizing textual features; similarly, minimizing $\mathcal{O}_2$ will push $\mathbf{v}_\ell$ exploit the most discriminative visual features.

[4]We use the software provided by http://koen.me/research/downloads/

Putting $\mathcal{O}_{main}$, $\mathcal{O}_1$ and $\mathcal{O}_2$ together, we thus have a new objective function for minimizing as:

$$\mathcal{J} = \mathcal{O}_{main} + \alpha_1\mathcal{O}_1 + \alpha_2\mathcal{O}_2 + \frac{\lambda_W}{2}\sum_{\ell\in\mathcal{L}}||\mathbf{w}_\ell||_2^2 + \frac{\lambda_V}{2}\sum_{\ell\in\mathcal{L}}||\mathbf{v}_\ell||_2^2 \tag{6}$$

where α_1 and α_2 are two parameters for balancing between the main objective and two subobjectives; $\frac{\lambda_W}{2}\sum_{\ell\in\mathcal{L}}||\mathbf{w}_\ell||_2^2$ and $\frac{\lambda_V}{2}\sum_{\ell\in\mathcal{L}}||\mathbf{v}_\ell||_2^2$ are two regularization terms to prevent from over-fitting; λ_W and λ_V are regularization parameters. We have tried ℓ_1-norm as regularization terms to make the solutions of $\{\mathbf{w}_\ell\}_{\ell\in\mathcal{L}}$ and $\{\mathbf{v}_\ell\}_{\ell\in\mathcal{L}}$ sparse, but it turns out the performance is not good. The reason may be the discriminative terms for POIs are usually infrequent textual (or visual) words. These words are forced to be zeros in the solution of $\{\mathbf{w}_\ell\}_{\ell\in\mathcal{L}}$ (or $\{\mathbf{v}_\ell\}_{\ell\in\mathcal{L}}$) when ℓ_1-norm is used, which thus deteriorates the performance.

When minimizing $\mathcal{J}$, the objective $\mathcal{O}_{main}$ is in charge of ranking the target POI correctly by combining textual context and visual content of photos, and the subobjectives $\mathcal{O}_1$ and $\mathcal{O}_2$ will push $\{\mathbf{w}_\ell\}_{\ell\in\mathcal{L}}$ and $\{\mathbf{v}_\ell\}_{\ell\in\mathcal{L}}$ to exploit the informative textual and visual features, respectively.

4.2.2 Incorporating User Preference

In this subsection, we discuss how to incorporate the user information of photos to make prediction. Intuitively, we aim at exploring users' visiting preferences over POIs, i.e., how likely a user visits a POI, which is useful for inferring where his/her photo is taken. According to **Observation** 3 in Section 3.2, however, users tend to take photos at previously unvisited POIs. Therefore, it is insufficient to model the visiting preference of a user based on his/her visited POIs. A good solution should be able to model the preferences of users over both visited and unvisited POIs. This makes exploiting users' visiting preferences in ranking POIs a non-trivial task.

To address the challenge, we propose a new approach based on factorization model to incorporate the user information. Factorization model has been proven to be a promising tool for capturing users' preferences, which has been applied in recommendation systems [10, 19, 15, 13]. As both users and POIs do not have explicit feature representations in our problem, we learn to embed users and POIs into a latent space for modeling their interactions. Specifically, we introduce a latent factor $\mathbf{u}_u$ of k-dimension for each user u and a latent factor $\mathbf{l}_\ell$ of k-dimension for each POI ℓ. By using them, we extend the scoring function in Eq. (2) as:

$$s_{up\ell} = \mathbf{w}_\ell \cdot \mathbf{x}_p + \mathbf{v}_\ell \cdot \mathbf{y}_p + \mathbf{u}_u \cdot \mathbf{l}_\ell \tag{7}$$

where $s_{up\ell}$ represents the interaction score among user u, photo p and POI ℓ. We incorporate the user preference in the ranking function by including the term $\mathbf{u}_u \cdot \mathbf{l}_\ell$. Different from textual and visual information, where $\mathbf{x}_p$ and $\mathbf{y}_p$ are predefined explicit representations in the terms $\mathbf{w}_\ell \cdot \mathbf{x}_p$ and $\mathbf{v}_\ell \cdot \mathbf{y}_p$, respectively, both $\mathbf{u}_u$ and $\mathbf{l}_\ell$ are unknown and require to be solved.

Replacing $s_{p\ell}$ with $s_{up\ell}$ in Eq. (3), the objective function can be extended as:

$$\mathcal{O}_{main'} = \sum_{\substack{p\in\mathcal{D},\\ \ell=p.\ell,u=p.u}} \sum_{\substack{\ell'\in\mathcal{L},\\ \ell'\neq\ell}} h(s_{up\ell} - s_{up\ell'}) \tag{8}$$

and the objective function $\mathcal{J}$ in Eq. (6) can also be updated by replacing $\mathcal{O}_{main}$ with $\mathcal{O}_{main'}$ here.

When minimizing the objective function $\mathcal{J}$, the latent factors of users and POIs will be optimized to encode users' visiting preferences. However, the optimization is a combined effect of textual, visual and user information, and the users' visiting preferences might not be effectively utilized. To address this issue, we design a subobjective for solely modeling the visiting preferences of users.

Our subobjective is performing matrix factorization on user-POI interaction data. Based on the training set $\mathcal{D}$, we can construct a $|\mathcal{U}| \times |\mathcal{L}|$ matrix $\mathbf{A}$, where $a_{u\ell}$ denote the frequency of user u takes photos at POI ℓ. Note that we use a transformation by setting $\tilde{a}_{u\ell} = \frac{1}{2}(\log(a_{u\ell}) + 1)$ in our implementation, because a small number of POIs may have very large frequency. Then we compute the latent factors of users and POIs, such that the user-POI interactions, modeled as inner product of users' latent factors and POIs' latent factors, are a good approximation of the matrix $\mathbf{A}$. However, approximating the matrix directly will over-highlight the users' preferences on visited POIs, because unvisited POIs have zero frequencies and directly approximating them will lead users' preferences over unvisited POIs close to zero, which is not reasonable for our problem since users tend to take photos at unvisited POIs and thus their preferences to unvisited POIs should not be too small. On the other hand, we cannot only approximate the non-zero entries in matrix $\mathbf{A}$, because the data is too sparse according to **Observation** 2 in Section 3.2. As a trade-off, we consider using weighted matrix factorization. In particular, we construct a $|\mathcal{U}| \times |\mathcal{L}|$ weighted matrix $\mathbf{B}$ by assigning larger weights for non-zero entries and smaller weights for zero entries as:

$$b_{u\ell} = \begin{cases} 1 & \text{if } a_{u\ell} > 0 \\ \varepsilon & otherwise \end{cases}$$

where $\varepsilon = 0.001$ is used in this paper. Because we give smaller weights for zero entries (corresponding to unvisited POIs), which is a relaxation for fitting zero entries, the users' preferences over unvisited POIs thus will not be very close to zero. The subobjective function of the task is written as:

$$\mathcal{O}_3 = \frac{1}{2}\sum_{u\in\mathcal{U}}\sum_{\ell\in\mathcal{L}} b_{u\ell}(\tilde{a}_{u\ell} - \mathbf{u}_u \cdot \mathbf{l}_l)^2 \tag{9}$$

The final objective function is thus changed into:

$$\mathcal{J}' = \mathcal{O}_{main'} + \alpha_1\mathcal{O}_1 + \alpha_2\mathcal{O}_2 + \alpha_3\mathcal{O}_3 + \frac{\lambda_W}{2}\sum_{\ell\in\mathcal{L}}||\mathbf{w}_\ell||_2^2 + \frac{\lambda_V}{2}\sum_{\ell\in\mathcal{L}}||\mathbf{v}_\ell||_2^2 + \frac{\lambda_{UL}}{2}\left(\sum_{u\in\mathcal{U}}||\mathbf{u}_u||_2^2 + \sum_{\ell\in\mathcal{L}}||\mathbf{l}_\ell||_2^2\right) \tag{10}$$

where α_3 is a parameter for controlling the importance of the user preference subtask; $\frac{\lambda_{UL}}{2}\left(\sum_{u\in\mathcal{U}}||\mathbf{u}_u||_2^2 + \sum_{\ell\in\mathcal{L}}||\mathbf{l}_\ell||_2^2\right)$ is a regularization term to prevent over-fitting problem and λ_{UL} is a regularization parameter. Here we use one regularization parameter for the latent factors of users and POIs, because only their inner product values matter to our objective function.

4.2.3 Optimization and Learning Algorithm

In this subsection, we discuss how we optimize the objective function $\mathcal{J}'$ to learn the prediction model. Let $\theta \in \{\mathbf{u}_u, \mathbf{l}_\ell, \mathbf{w}_\ell, \mathbf{v}_\ell\}$, where $u \in \mathcal{U}$ and $\ell \in \mathcal{L}$, denote the param-

eters of our model. We adopt the Stochastic Gradient Descent (SGD) method for learning them. As $\mathcal{J}'$ is composed of one main objective $\mathcal{O}_{main'}$ and three subobjectives $\alpha_1\mathcal{O}_1$, $\alpha_2\mathcal{O}_2$ and $\alpha_3\mathcal{O}_3$, we perform the SGD updates alternatively for these objectives. That is, in each iteration, we sample one training instance $p \in \mathcal{D}$ and then update parameters as follows for each objective:

$$\textit{main objective}: \quad \theta \leftarrow \theta - \gamma\frac{\partial \mathcal{O}_{main'}}{\partial \theta} \tag{11}$$

$$\textit{subobjective 1}: \quad \theta \leftarrow \theta - \gamma\frac{\partial \alpha_1\mathcal{O}_1}{\partial \theta} \tag{12}$$

$$\textit{subobjective 2}: \quad \theta \leftarrow \theta - \gamma\frac{\partial \alpha_2\mathcal{O}_2}{\partial \theta} \tag{13}$$

$$\textit{subobjective 3}: \quad \theta \leftarrow \theta - \gamma\frac{\partial \alpha_3\mathcal{O}_3}{\partial \theta} \tag{14}$$

where γ is the learning rate.

Next, we show how to calculate the corresponding gradients used in above updating formulae. As the main objective, subobjective 1 and subobjective 2 are similar, we derive the gradients for the main objective here as an example:

$$-\frac{\partial \mathcal{O}_{main'}}{\partial \theta} = \begin{cases} \mathbf{l}_\ell - \mathbf{l}_{\ell'} - \lambda_{UL}\mathbf{u}_u & \text{if } \theta = \mathbf{u}_u, \\ \mathbf{u}_u - \lambda_{UL}\mathbf{l}_\ell & \text{if } \theta = \mathbf{l}_\ell, \\ -\mathbf{u}_u - \lambda_{UL}\mathbf{l}_{\ell'} & \text{if } \theta = \mathbf{l}_{\ell'}, \\ \mathbf{x}_p - \lambda_W\mathbf{w}_\ell & \text{if } \theta = \mathbf{w}_\ell, \\ -\mathbf{x}_p - \lambda_W\mathbf{w}_{\ell'} & \text{if } \theta = \mathbf{w}_{\ell'}, \\ \mathbf{y}_p - \lambda_V\mathbf{v}_\ell & \text{if } \theta = \mathbf{v}_\ell, \\ -\mathbf{y}_p - \lambda_V\mathbf{v}_{\ell'} & \text{if } \theta = \mathbf{v}_{\ell'}, \\ 0 & \textit{otherwise}. \end{cases}$$

We note that only if $s_{up\ell} - s_{up\ell'} < 1$, the hinge-loss is nonzero and we use the above gradient calculations for updating parameters; otherwise, the hinge-loss is zero and we do not need to update parameters. Moreover, the corresponding regularization terms are incorporated into the calculations. For subobjectives 1 and 2, $\frac{\partial \alpha_1\mathcal{O}_1}{\partial \theta}$ and $\frac{\partial \alpha_2\mathcal{O}_2}{\partial \theta}$ can be computed in the similar way. The gradients for subobjective 3 are derived as follows:

$$-\frac{\partial \alpha_3\mathcal{O}_3}{\partial \theta} = \begin{cases} \alpha_3 \sum\limits_{\ell \in \mathcal{L}} b_{u\ell}(a_{u\ell} - \mathbf{u}_u \cdot \mathbf{l}_\ell)\mathbf{l}_\ell - \alpha_3\lambda_{UL}\mathbf{u}_u & \text{if } \theta = \mathbf{u}_u, \\ \alpha_3 \sum\limits_{u \in \mathcal{U}} b_{u\ell}(a_{u\ell} - \mathbf{u}_u \cdot \mathbf{l}_\ell)\mathbf{u}_u - \alpha_3\lambda_{UL}\mathbf{l}_\ell & \text{if } \theta = \mathbf{l}_\ell, \\ 0 & \textit{otherwise}. \end{cases}$$

We note that when deriving the gradients for each subobjective, the corresponding regularization terms should be incorporated.

The proposed method, Ranking with Textual, Visual and User information for Photos to POIs, called RankTVU_P2P, can thus be summarized as Algorithm 1. In the algorithm, we iteratively update the model parameters based on the main objective and three subobjectives until the performance on validation set is stable and does not increase. After obtaining the model parameters, the mapping score of a photo to a POI is calculated as Eq. (7). The larger the score is, the better the photo is mapped to the corresponding POI.

Algorithm 1: RankTVU_P2P

input : training set $\mathcal{D}$, validation set $\mathcal{V}$, learning rate γ, parameters α_1, α_2 and α_3
output: model parameters θ

Initialize θ with Normal distribution $\mathcal{N}(0, 0.01)$;
repeat
  for $p \in \mathcal{D}$ **do**
    $\theta \leftarrow \theta - \gamma\frac{\partial \mathcal{O}_{main'}}{\partial \theta}$;
    $\theta \leftarrow \theta - \gamma\frac{\partial \alpha_1\mathcal{O}_1}{\partial \theta}$;
    $\theta \leftarrow \theta - \gamma\frac{\partial \alpha_2\mathcal{O}_2}{\partial \theta}$;
    $\theta \leftarrow \theta - \gamma\frac{\partial \alpha_3\mathcal{O}_3}{\partial \theta}$;
until *the performance on validation set* $\mathcal{V}$ *does not increase*;
return θ

5. EXPERIMENTS

In this section, we conduct extensive experiments to evaluate our proposed technique with the existing three state-of-the-arts, based on two Instagram datasets that we have crawled from the users in New York City (NYC) and Singapore (SG), respectively.

5.1 Experimental Setup

5.1.1 Data Preprocessing

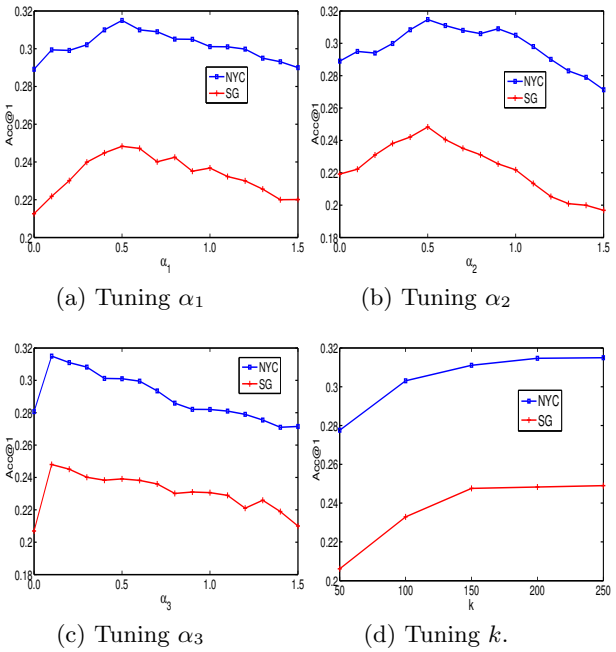
For NYC and SG datasets we crawled, we use the most recent three months data, i.e., from 13 Aug 2014 to 13 Nov 2014, in our experiments to evaluate the performance of different algorithms. For data preprocessing, we remove POIs with less than 5 photos and users who visited less than 10 and 5 POIs in NYC and SG, respectively.[5] As our aim is to test the performance of photo mapping, we only use the photos that have already been associated with POIs by Instagram users. We construct the textual representation of photos by keeping the words with frequency larger than 10 into textual vocabulary. Note that we employ the word frequency in the bag-of-word model as discussed in Section 4.1. For the visual representation, we construct 1000 visual words as vocabulary by using the method introduced in Section 4.1. After preprocessing, NYC dataset comprises 2,049 POIs and 74,758 photos from 3,556 users, and SG dataset comprises 1,363 POIs and 58,072 photos from 5,717 users. Each photo is represented as a 1000-dimensional visual feature vector, and represented as a 8217-dimensional and a 6696-dimensional textual feature vector for NYC and SG, respectively. Then, for each user, we mark off 20% of his/her most recent photos to build the test set and mark off another 10% earlier photos as tuning/validation set. The remaining (the earliest) 70% photos are employed as training set for building the prediction models. The datasets are available at http://www.ntu.edu.sg/home/gaocong/data/Instagram.zip

5.1.2 Evaluation Metrics

We employ two standard metrics to evaluate the performance of photo associating results as in [20], namely, Mean Reciprocal Rank and Accuracy within top N, denoted by MRR and Acc@N, respectively, where N is the number of candidate POIs produced by algorithms. MRR is the average of the reciprocal ranks of the results produced for all the photos in test set. Specifically, it is computed as follows:

$$MRR = \frac{1}{|\mathcal{P}_{test}|}\sum_{p \in \mathcal{P}_{test}}\frac{1}{rank_{p.\ell}} \tag{15}$$

[5]We perform the preprocessing to reduce noises. We note that after the preprocessing the data is still very sparse, where 51.7% and 47.9% POIs have less than 10 photos in NYC and SG, respectively.



(a) Tuning α_1 (b) Tuning α_2

(c) Tuning α_3 (d) Tuning k.

Figure 6: Parameter tuning.

where $rank_{p.\ell}$ is the position of target POI $p.\ell$ in the ranking list produced by algorithms for the test photo p; $\mathcal{P}_{test}$ denotes the test set. Clearly, a bigger MRR indicates the target POI is highly ranked and thus a better result. Acc@N is computed as:

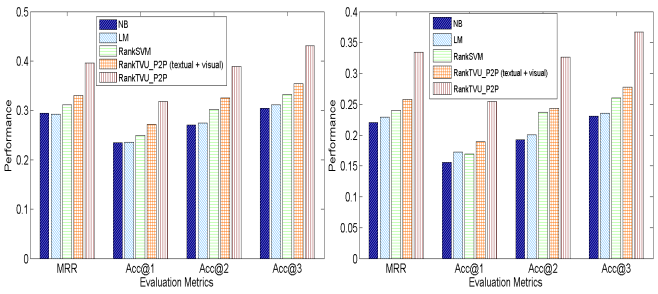
$$Acc@N = \frac{1}{|\mathcal{P}_{test}|} \sum_{p \in \mathcal{P}_{test}} |Top(N,p) \cap \{p.\ell\}| \qquad (16)$$

where $Top(N,p)$ denotes the set of top-N candidate POIs produced by algorithms for the test photo p. Acc@N considers a mapping is correct as long as the target POI $p.\ell$ for photo p is ranked within top-N positions. For N, we use the values of 1, 2 and 3 ($N = 1$ is the default value) in the experiments, respectively, as the accuracies of those top predicted results are definitely more important for our problem.

5.1.3 Baseline Algorithms

The following three related methods are utilized as the baselines.

- NB: This is the Naive Bayesian method, which has been used in the landmark identification task [5]. We build an NB model for textual and visual content individually, and subsequently use the optimal linear combination of the two models to make a prediction.
- LM: This is the Language Model. As shown in Section 2, it is proposed for placing photos onto grids [20, 16]. Similarly, for combining visual and textual information, we build a LM for each of them, and then adopt a linear combination of both models for associating photos with POIs.
- RankSVM: As our proposed method is ranking based method, we also employ state-of-the-art ranking technique RankSVM [9] as a baseline.



(a) NYC (b) SG

Figure 7: Performance comparisons.

All the existing methods cannot leverage user information to infer users' visiting preferences to unvisited POIs. Different from them, our approach has a weighted matrix factorization subobjective $\mathcal{O}_3$, which acts as a recommender component and is capable of exploiting such information. For a fair comparison, we use our tuning set to find the optimal parameters for all the baseline methods, and then use them to evaluate the performance on test set.

5.2 Experimental Results

5.2.1 Parameter Setting and Tuning

In the experiments, we set the regularization parameters $\lambda_V = \lambda_{UL} = 0.1$ and $\lambda_W = 0.001$, and set the learning rate γ as a small value 0.01 in our experiments to ensure the generalization accuracy. Next, we tune the parameters α_1, α_2 and α_3 used in three subobjectives, and the dimension k of latent factors of users and POIs, based on tuning set, to show how they affect the performance of the proposed method. For parameters α_1, α_2 and α_3, we first perform grid search to find the optimal combination of them, and then tune each parameter by fixing the other twos to demonstrate their individual impact on the performance.

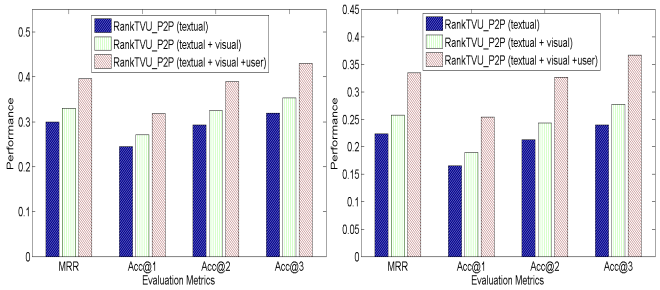
Figure 6(a) shows how the performance of the proposed method changes when we tune the parameter α_1, which controls the weight of textual subojective $\mathcal{O}_1$. We can see from this figure the proposed method performs the best at $\alpha_1 = 0.5$ on both NYC and SG datasets. We show the effect of parameter α_2, controlling the weight of visual subobjective $\mathcal{O}_2$, in the Figure 6(b). It can be seen that the best performance is delivered at $\alpha_2 = 0.5$ on the two datasets. Figure 6(c) shows how the parameter α_3, used for combining user preference subobjective $\mathcal{O}_3$, affects the performance of the proposed method. We observe that the performance first increases and then decreases as α_3 is increased, and the best performance is yielded at $\alpha_3 = 0.1$ for both datasets. We note that although α_1 and α_2 are larger than α_3, we cannot conclude that the textual and visual subobjectives are more important than user preference subobjective, because the subojective functions $\mathcal{O}_1$, $\mathcal{O}_2$ and $\mathcal{O}_3$ may be in different scales. We will analyze how the three subobjectives affect the performance in the next subsection. Finally, we show the effect of dimension k of latent factors on performance in Figure 6(d). We find that the performance increases as the dimension k is increased. After k exceeding 200, the performance does not change significantly and thus we use k=200 in our experiments.

5.2.2 *Performance Comparison and Analysis*

Performance Comparison. Figure 7 presents the performance comparison results of different algorithms. RankTVP_P2P (textual + visual) denotes our approach utilizing only textual and visual content. We can see from the figure that the proposed method RankTVU_P2P performs the best, in terms of two evaluation metrics, namely MRR and Acc@N. It outperforms RankSVM, LM and NB, in terms of Acc@1, by more than 27.25% and 41.75% on NYC and SG, respectively. This is because RankTVU_P2P exploits the users' visiting preferences over POIs (but RankSVM, LM and NB cannot leverage), which is an important information source for handling the sparsity problem of training data. When only utilizing the textual and visual content, we observe that our approach still delivers better performance than the baseline algorithms. In terms of Acc@1, the improvements are more than 12.5% and 11.7% on the two datasets, respectively. Moreover, we can see that RankSVM outperforms the other two baseline methods, which are all classifier based approaches. This may be attributed to that RankSVM, as a ranking based method, is more suitable to address the mapping problem when training samples are scarce. Finally, we observe that the Acc@N increases as N increases for all the methods. This is because the evaluation metric Acc@N considers a mapping is correct as long as the target POI for a test photo is ranked within top-N positions. We find that the proposed RankTVU_P2P method consistently performs better than the baselines for different values of N.

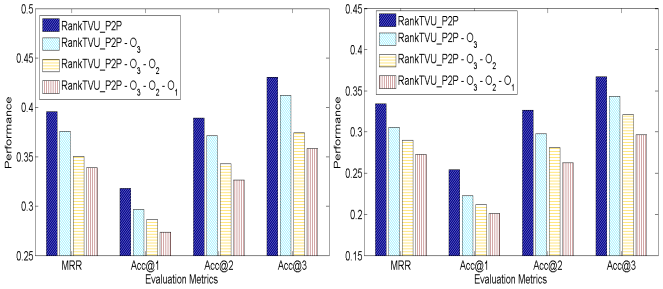
Impact of Different Types of Information. Next, we analyze how different types of information affect our results. We incorporate the information one by one into our proposed RankTVU_P2P to test how the performance changes. Figure 8 shows the detailed results on both datasets. We observe the performance of our model increases when textual, visual and user information are incorporated one by one in terms of both MRR and Acc@N. When combing textual context with visual content, the performance improves, in terms of Acc@1, 11.0% and 14.1% on NYC and SG, respectively. When we further incorporate the user information, in terms of Acc@1 the performance boost 17.2% and 34.4% on both datasets, respectively. Finally, we find our model with only textual context achieves 76.9% and 65.2% of the best performance with all the three types of information incorporated, on NYC and SG, respectively. The observations demonstrate the usefulness of each type of information. As the three types of information are from different views, they may contain complementary information to enhance each other for identifying the target POIs, which is a key reason that we obtain substantial improvements when combining them. According to the results, we also find that although visual content is important for associating photos with POIs, textual and user information have greater contributions to the final results.

Impact of Subobjectives. Finally, we study how our designed subobjectives in RankTVU_P2P help us obtain a better performance. To determine their individual effects on our proposed RankTVU_P2P performance, we take apart the subobjective functions from RankTVU_P2P one by one, and check their impacts on the performance. Figure 9 presents the results on both datasets, where we use the suffix "$-\mathcal{O}_i$" to denote the result obtained by removing subjective $\mathcal{O}_i$ (i=1, 2 and 3) from RankTVU_P2P.



(a) NYC (b) SG

Figure 8: Impact of different types of information.



(a) NYC (b) SG

Figure 9: Impact of subobjectives.

We observe that the performance of RankTVU_P2P decreases when we remove the subobjectives $\mathcal{O}_3$, $\mathcal{O}_2$ and $\mathcal{O}_1$ one by one. Comparing RankTVU_P2P with RankTVU_P2P-$\mathcal{O}_3$, we find that incorporating the subobjective $\mathcal{O}_3$ improves the performance, in terms of Acc@1, 7.2% and 14.2% on NYC and SG, respectively. Moreover, by checking the results for RankTVU_P2P-$\mathcal{O}_3$ and RankTVU_P2P-$\mathcal{O}_3$-$\mathcal{O}_2$, we can see that 3.6% and 5.1% improvements, in terms of Acc@1, are obtained with the incorporation of the subobjective $\mathcal{O}_2$ on both datasets, respectively. Finally, studying the results of RankTVU_P2P-$\mathcal{O}_3$-$\mathcal{O}_2$ and RankTVU_P2P-$\mathcal{O}_3$-$\mathcal{O}_2$-$\mathcal{O}_1$, we observe that the subobjective $\mathcal{O}_1$ leads to 4.7% and 5.5% improvements in terms of Acc@1 on NYC and SG, respectively.

The results demonstrate these three designed subobjectives play crucial roles for helping RankTVU_P2P obtain better performance. The reason is that the subobjectives ($\mathcal{O}_1$, $\mathcal{O}_2$ and $\mathcal{O}_3$) will push the scoring function for POIs to exploit the most useful information from textual context, visual content and user information of photos.

6. CONCLUSIONS

In this paper, we define and study the problem of associating Instagram photos with POIs. To address the problem, we first find some interesting patterns contained in Instagram by data analysis, and then propose a ranking based mapping method, called RankTVU_P2P. In the proposed method, a new multi-task objective function is developed, where the main objective is to score POIs with the three types of information, namely, textual context, visual content, and user, and the other subobjectives are to maximize

the prediction effectiveness for each type of information. In particular, we design a weighted matrix factorization subobjective to learn the visiting preferences of users over POIs. A stochastic gradient descent based algorithm is developed to optimize the objective function for learning parameters of our model. Extensive experiments are conducted on two sets of Instagram data, namely, NYC and SG, and the results show that our model outperforms the baseline methods substantially.

Although Instagram has gained increasing popularity, little research has been conducted on it [8, 23, 7]. We propose a new research problem based on Instagram in this paper. Based on our work, several interesting problems can be investigated or studied in the future. First, we only consider the Instagram photos that can be associated with POIs in our work such that we just focus on how to solve the mapping problem. However, it would be interesting to study whether a photo from Instagram is really related to POIs or not in the future. Second, there are other types of information that could be investigated in Instagram, e.g., time stamps of photos, coordinates of photos (only available for the photos taken when GPS is enabled) and category information of POIs, etc.

7. ACKNOWLEDGEMENT

This work is supported in part by a grant awarded by a Singapore MOE AcRF Tier 2 Grant (ARC30/12) and a grant awarded by Microsoft Research Asia.

8. REFERENCES

[1] Y. Avrithis, Y. Kalantidis, G. Tolias, and E. Spyrou. Retrieving landmark and non-landmark images from community photo collections. In *Proceedings of ICM*, pages 153–162. ACM, 2010.

[2] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of WWW*, pages 61–70. ACM, 2010.

[3] W.-C. Chen, A. Battestini, N. Gelfand, and V. Setlur. Visual summaries of popular landmarks from community photo collections. In *The Forty-Third Asilomar Conference on Signals, Systems and Computers*, pages 1248–1255. IEEE, 2009.

[4] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of CIKM*, pages 759–768. ACM, 2010.

[5] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of WWW*, pages 761–770. ACM, 2009.

[6] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *CVPR*, pages 1–8. IEEE, 2008.

[7] N. Hochman and L. Manovich. Zooming into an instagram city: Reading the local through social media. *First Monday*, 18(7), 2013.

[8] Y. Hu, L. Manikonda, S. Kambhampati, et al. What we instagram: A first analysis of instagram photo content and user types. In *Proceedings of ICWSM*, AAAI, 2014.

[9] T. Joachims. Training linear SVMs in linear time. In *Proceedings of SIGKDD*, pages 217–226. ACM, 2006.

[10] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[11] D. Leung and S. Newsam. Proximate sensing: Inferring what-is-where from georeferenced photo collections. In *CVPR*, pages 2955–2962. IEEE, 2010.

[12] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of SIGKDD*, pages 1023–1031. ACM, 2012.

[13] X. Li, G. Cong, X. Li, T. A. N. Pham, and S. Krishnaswamy. Rank-GeoFM: A ranking based geographical factorization method for point of interest recommendation. In *Proceedings of SIGIR*, ACM, 2015.

[14] Y. Li, D. J. Crandall, and D. P. Huttenlocher. Landmark classification in large-scale image collections. In *ICCV*, pages 1957–1964. IEEE, 2009.

[15] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, and Y. Rui. GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation. In *Proceedings of SIGKDD*, pages 831–840. ACM, 2014.

[16] B. Liu, Q. Yuan, G. Cong, and D. Xu. Where your photo is taken: Geolocation prediction for social images. *JASIST*, 65(6):1232–1243, 2014.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[18] F. O. Ostermann, M. Tomko, and R. Purves. User evaluation of automatically generated keywords and toponyms for geo-referenced images. *JASIST*, 64(3):480–499, 2013.

[19] S. Rendle. Factorization machines. In *ICDM*, pages 995–1000. IEEE, 2010.

[20] P. Serdyukov, V. Murdock, and R. Van Zwol. Placing flickr photos on a map. In *Proceedings of SIGIR*, pages 484–491. ACM, 2009.

[21] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove. Mobile call graphs: beyond power-law and lognormal distributions. In *Proceedings of SIGKDD*, pages 596–604. ACM, 2008.

[22] T. H. Silva, P. O. Melo, J. M. Almeida, J. Salles, and A. A. Loureiro. A picture of instagram is worth more than a thousand words: Workload characterization and application. In *DCOSS*, pages 123–132. IEEE, 2013.

[23] T. H. Silva, P. O. Vaz de Melo, J. M. Almeida, J. Salles, and A. A. Loureiro. A comparison of foursquare and instagram to the study of city dynamics and urban social behavior. In *Proceedings of SIGKDD International Workshop on Urban Computing*, page 4. ACM, 2013.

[24] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann. Who, where, when and what: discover spatio-temporal topics for twitter users. In *Proceedings of SIGKDD*, pages 605–613. ACM, 2013.

[25] Y.-T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. Tour the world: building a web-scale landmark recognition engine. In *CVPR*, pages 1085–1092. IEEE, 2009.