Location-Sensitive User Profiling Using Crowdsourced Labels

Wei Niu, James Caverlee, Haokai Lu
Department of Computer Science and Engineering, Texas A&M University
College Station, TX 77840, USA
{wei,caverlee,hlu}@cse.tamu.edu

Abstract

In this paper, we investigate the impact of spatial variation on the construction of *location-sensitive user profiles*. We demonstrate evidence of spatial variation over a collection of Twitter Lists, wherein we find that crowdsourced labels are constrained by distance. For example, that `energy` in San Francisco is more associated with the `green movement`, whereas in Houston it is more associated with `oil and gas`. We propose a three-step framework for location-sensitive user profiling: first, it constructs a crowdsourced label similarity graph, where each labeler and labelee are annotated with a geographic coordinate; second, it transforms this similarity graph into a directed weighted tree that imposes a hierarchical structure over these labels; third, it embeds this location-sensitive folksonomy into a user profile ranking algorithm that outputs a ranked list of candidate labels for a partially observed user profile. Through extensive experiments over a Twitter list dataset, we demonstrate the effectiveness of this location-sensitive user profiling.

Introduction

User profiles are a valuable component of many applications, including recommender systems, search engines, question-answering systems, and online social networks. These profiles provide insight into the interests and expertise of each user, and can lead to improved personalization of the underlying system (Liu et al. 2012a; Majumder and Shrivastava 2013; Weng et al. 2010). Many systems rely on an *explicit* definition of a user profile – for example, by filling in an "About" section in a social media profile or by directly selecting topics of interest on a question-answer system. Alternatively, *implicit* user profiles can be uncovered through methods like query log mining, running Latent Dirichlet Allocation (LDA) over a user's posts, or by applying matrix factorization approaches to identify hidden (or latent) topics of interest (Hong, Doumith, and Davison 2013; Jiang et al. 2012; Yin et al. 2014; Zhong et al. 2015). In a complementary direction, recent years have seen the development of *crowdsourced* user profiles construction, e.g., (Bhattacharya et al. 2014; Ghosh et al. 2012; Rakesh et al. 2014). In this scenario, crowds of users apply descriptive labels on other users, so that in the aggregate these labels provide a crowdsourced user profile of the target user. For example, Twitter Lists and LinkedIn's Skill tags provide partial perspective on what users are known for by aggregating crowd labeling knowledge. However, the vast majority of users have no crowd labels; their expertise are essentially hidden from important applications such as personalized recommendation, community detection, and expert mining.

In this paper, we aim to extend the reach of these crowdsourced methods, so as to construct robust user profiles for the long-tail of users for whom we have incomplete labels. A natural approach to extend the reach of these crowd-generated labels is to apply existing *tag recommendation* methods (Brooks and Montanez 2006; Heymann, Ramage, and Garcia-Molina 2008; Sigurbjornsson and Van Zwol 2008; Tuarob, Pouchard, and Giles 2013; Xia et al. 2013). However, many of these approaches have viewed tag relationships without regard for the local variations that are inherent in real-world crowdsourced labels of users. For example, we find in a sample of Twitter Lists that the label `energy` in San Francisco is more associated with the `green movement`, whereas in Houston it is more associated with `oil and gas`. These spatial variations are a critical component of crowdsourced labels and require careful consideration. Beyond just the presence of different relationships across locations, there is often a variation in the strength of this relationship from location to location. For example, `stock` and `finance` are more closely related in New York City than in Portland. Further, there is even a potential for varying location-specific senses of a tag (polysemy). For example, the tag `rockets` in the Houston area may be associated with the local NBA team instead of other senses of the word.

Hence, we explore the impact of spatial variation on the construction of *location-sensitive user profiles*. Our main intuition is that spatial variation over crowdsourced labels can be modeled in a location-sensitive folksonomy to provide a comprehensive and up-to-date picture of location-aware topics, topic relations, and a fine-grained topic level view of the social media corpus, which may mitigate the sparsity inherent in the raw labels. Recent studies (Zhu et al. 2015; Wang et al. 2015) have shown how hierarchical topic structures can improve ranking and recommendation, indicating the importance of folksonomies. Thus, we aim to study the impact of location-sensitive hierarchical structures of

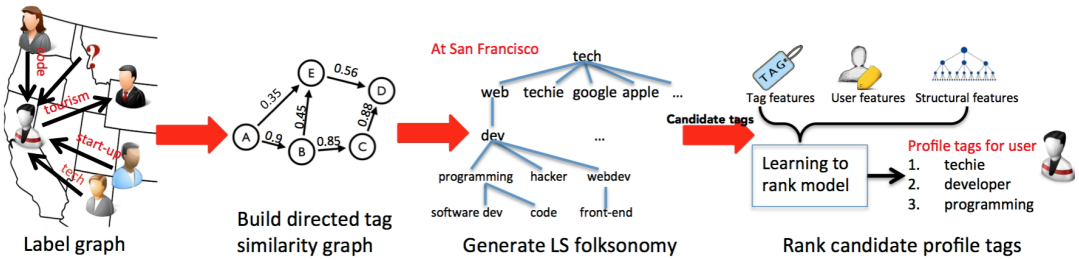

Figure 1: Overall approach: constructing location-sensitive user profiles from crowdsourced labels

crowdsourced tags on user profiling.

Concretely, we propose an approach for *location-sensitive user profiling* as illustrated in Figure 1. First, we construct a crowdsourced label similarity graph induced from crowdsourced labels, where each labeler and labelee are annotated with a geographic coordinate; this similarity graph varies by location to capture spatial variations of the kind identified above (e.g., the similarity graph for San Francisco will link `energy` with `green movement`). Second, we transform this similarity graph into a directed weighted tree that imposes a hierarchical structure over these labels, such that labels like `sports` are higher in the tree, whereas labels like `rockets` are lower, thereby providing finer granularity for building user profiles. Finally, we embed this location-sensitive folksonomy into a user profile ranking algorithm that outputs a ranked list of candidate labels for a partially observed user profile. Through extensive experiments over a Twitter list dataset, we demonstrate the effectiveness of this location-sensitive user profile estimation.

Related Work

In this section, we highlight several research directions that inform the work presented here.

User and Resource Profiling. User profiling is critical for enabling effective information services. Many efforts are devoted to profile a user's topic interest for applications in personalized search (Qiu and Cho 2006), targeted advertisement (Ahmed et al. 2011), social media (Zhao et al. 2015), and expert search (Ribeiro et al. 2015). Many other research efforts aim to reveal users' demographic information like age and gender (Ikeda et al. 2013; Li, Ritter, and Hovy 2014; Li, Wang, and Chang 2014). For example, Li et al. (Li, Wang, and Chang 2014) propose a co-profiling approach to profile users' attributes like employer, college, and circles of friends in a joint fashion.

Meanwhile, another line of research focuses on recommending tags to resources, targeting the sparsity of collaborative tagging in order to construct comprehensive tag profiles (Sigurbjornsson and Van Zwol 2008; Tuarob, Pouchard, and Giles 2013; Xia et al. 2013). Some proposed approaches are based on topic models, association rule mining, and context information (Heymann, Ramage, and Garcia-Molina 2008; Krestel, Fankhauser, and Nejdl 2009; Tuarob, Pouchard, and Giles 2013). Later works construct hierarchical folkonomies to assist resource tag prediction (Song, Qiu, and Farooq 2011; Verma et al. 2015). In our work, we leverage location-sensitive folksonomy for user profiling and we adopt a learning to rank approach that automatically weighs a group of factors to minimize the prediction error.

Folksonomy Construction. Considerable research has been devoted to folksonomy generation (Heymann and Garcia-Molina 2006; Plangprasopchok and Lerman 2009; Schmitz 2006; Rego, Marinho, and Pires 2015; Brooks and Montanez 2006; Liu et al. 2012b). Many approaches consider the co-occurrence of tag pairs (Heymann and Garcia-Molina 2006; Schmitz 2006; Song, Qiu, and Farooq 2011). Generally, they first identify subsumption relations of tag pairs using unsupervised approaches and then prune these relations into a tree. Our method differs in two key aspects: (i) we are constructing a location-sensitive foklsonomy that models the knowledge structures focused on a particular location; and (ii) we propose to induce the folksonomy using an optimization algorithm that best conserves the graph structure.

Geographic Influence. The impact of geographical distance has been widely studied for online social interactions (Kaltenbrunner, Scellato, and others 2012; Scellato et al. 2011). Additionally, there are also studies of spatial variation over query logs and social media, e.g., (Backstrom et al. 2008; Brodersen, Scellato, and Wattenhofer 2012; Cheng, Caverlee, and Lee 2010; Hu, Sun, and Liu 2014; Zhang et al. 2012). Our intuition is that due to the geographic, cultural, and structural differences among locations, there could be corresponding differences reflecting how people organize information in these locations.

Location-Sensitive User Profiling

Our overarching goal is to estimate high-quality user profiles that respect this observed spatial variation. We assume some partial coverage of users via existing crowdsourced tags (e.g., from Twitter Lists or LinkedIn's Skill Tags), but that many tags are unknown. That is, given a user u's full (but hidden) tag profile $P(u)$, we have visibility only to some portion of this profile $P_k(u)$ where $P_k(u) \subset P(u)$. The goal is to estimate the unseen tags t_i of u where $t_i \in P(u) - P_k(u)$. Our intuition is that the spatial variation of how tags are applied can be carefully modeled to create high-quality user profiles.

Spatial Variation in Crowdsourced Labels

In this section, we provide data-driven evidence for spatial variation in crowdsourced labels from a collection of Twitter

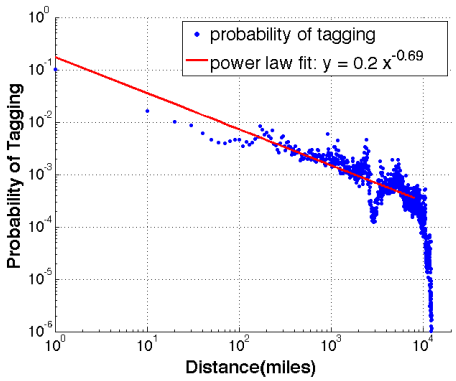

Figure 2: Probability of tagging as a function of distance between labeler and user.

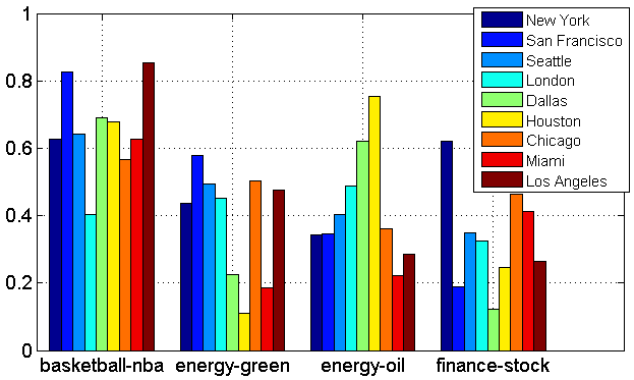

Figure 3: Example Tag Pairs Similarity.

lists (described more fully in Experiments). Twitter lists are one form of crowdsourced tagging, whereby individual users can add other users to a curated list annotated with a name.

How does distance impact tagging?
We begin by investigating in Figure 2 the impact of distance on the probability that a list labeler will include another user on a list. We observe that the probability of tagging is exponentially decaying with distance, which indicates a user is less likely to be tagged by a labeler as the distance between them increases. This spatial locality is a well-known property of many offline relationships and has been confirmed repeatedly even in online scenarios where distance is not inherently a limiting factor. This locality of tagging suggests that a method for user profile prediction that is induced from these crowd-based tags should reflect local knowledge; that is, since tags are not uniformly applied across distances, there may exists local variations of interest.

Example location-sensitive relationships.
We consider the relationships between pairs of labels across different locations in our dataset. Representing each tag as a location-specific vector (see the following section for additional details), we show in Figure 3 the relationships of a group of tag pairs at multiple locations using cosine similarity. The x-axis shows each tag pair and each color bar represents the similarity at a location. We observe that the magnitude of tag-pair relations varies across different locations. For example, we find the similarity between general concepts like `nba` and `basketball` tends to be relatively even across locations, with London having the lowest value; `finance` and `stock` have highest confidence in New York while lowest in Houston. Another typical example is the similarity between `energy` with `green` and `oil`. Interestingly, we notice `energy` and `oil` have the strongest relationship in Houston, while `energy` and `green` bond closest in San Francisco. This fits our understanding of these locations, since Texas is a major oil and gas hub, while San Francisco is a more eco-friendly community. These phenomena suggest location-sensitive user profiling has potential to reflect the characteristics of these locations.

Next, we shed light on the three-step framework (as shown in Figure 1).

Crowdsourced Label Similarity Graph

Given a group of users $\mathcal{U} = \{u_1, u_2, ..., u_n\}$, where each user u is associated with a geographic coordinate l_u and a tag profile P_u which contains a variable number of (tag, frequency) pairs $\{(t_1, f_1), (t_2, f_2), ...\}$. Our goal is to build a location-sensitive tag similarity graph from these profiles.

We begin by proposing a distance weighting scheme which weights the profile tags of a user according to how far this user is from the target location. Our intuition is a distant labeler is considered less knowledgeable about local users. We adopt a model popularized in the GIS literature – the *zone of indifference model* – for capturing this spatial influence. The key idea is to combine the inverse distance with a fixed distance band model. In this model, all users within the distance band are considered equally important and once beyond the threshold distance, a user's influence drops off quickly following an exponential rate. We empirically set the distance band as 50 miles for large cities, which defines a circle area centered at the target location. Hence, the weight of a user w.r.t the target location l_t is

$$w_u(l_t) = \begin{cases} 1 & \text{if } d \leq 50 \\ (\frac{d-50}{50})^{-\alpha} & \text{else} \end{cases}$$

where d is the distance from a user's location l_u to l_t and α is a constant, set experimentally. Thus, we utilize the whole dataset for constructing the location-sensitive folksonomy for each location. This avoids the sparsity issues that may arise (if we were to build a location-sensitive folksonomy using only locally-available tags) and mitigates data imbalances across locations (so a smaller city is not penalized in folksonomy creation relative to a larger city). We represent each tag t_i at target location l_t as a tf-idf vector of the users who are labeled with the tag, and each user is weighted by her corresponding influence $w_u(l_t)$: $\boldsymbol{t_i'} = \boldsymbol{t_i} \cdot \boldsymbol{w_i}$. Tag vectors vary in each location according to the users and weights.

With these tag vectors $\boldsymbol{t_i'}$, we then construct a directed tag similarity graph. We compare four similarity measures for tag pairs. Three are symmetric measures including cosine, RBF kernel, and pointwise mutual information (PMI). The

fourth measure is imbalanced (meaning the strength of one tag to a second tag is not necessarily the same as in the reverse situation) and based on a modified version of the traditional association rules notion of confidence (what we term *modified confidence*, or MConf).

Modified confidence. Finally, we adopt the idea of confidence from association rule mining which has been used for inferring subsumption relations between tags. Since instead of predicting very general tags, we would like to predict tags that are as specific as possible. Thus we propose a modified confidence metric $(t_i \Rightarrow t_j)$:

$$MConf(t_i \Rightarrow t_j) = \frac{C(t_i,t_j)}{F(t_i)} \cdot (1 - \frac{|\log(F(t_j)) - \log(F(t_i))|}{\log(\max F)})$$

where $C(t_i,t_j) = \min(f_1, f_2)$ is the co-occurrence frequency of the tag pair, $F(t_i) = \sum_u f_i$ is the overall frequency of t_i in the corpus and $\max F$ is the overall frequency of the most used tag in the corpus. Here we use confidence as a criteria for ordering tags and we only consider confidence for cases that $conf(t_i \Rightarrow t_j) \geq conf(t_j \Rightarrow t_i)$. The reason of multiplying a weight is to avoid connecting tags with large frequency difference as mentioned in the second drawback. Even if confidence is high, there might be some intermediate nodes that fit between the nodes. As tag frequency versus number of tags follow a power law distribution, we model frequency difference with damping factor which is a fraction in log scale. When there is no frequency difference, the factor is 1, when there is large frequency difference, the factor decays to 0.

Location-Sensitive Folksonomy Construction

Given these measures of tag similarity that capture both user and location influences and similarity graph, we next turn to induce a folksonomy bound to a target location l, which is represented as directed rooted tree (arborescence) $\mathcal{T} = (\mathcal{V}, \mathcal{E})$, where each tag has a unique parent. The node set $\mathcal{V}$ contains all unique tags that $\mathcal{U}$ has been labeled with, and edge set $\mathcal{E}$ contains subsumption relations of tag pairs. The abstractness of each tag is controlled by its level in the tree. We can further assign weights to edges to capture the similarity between tag pairs. Note that we will build a different location-sensitive folksonomy for each location of interest (e.g., one for Houston, one for Chicago).

In order to define an order of abstractness for the tags, we calculate the closeness centrality of each tag, defined as: $Centrality(t) = \sum_j \mathrm{sim(t, t_j)}$, which sums up each tag's similarity with all other tags. This definition forces general tags to have high centrality. Since the modified confidence metric already assigns a direction for a pair of tags, this step is exempted. Then we organize the tags and relations into a directed weighted graph $\mathcal{G}$. To do this, we initialize $\mathcal{G}$ with a ROOT node and add an edge for the tag pair when the similarity is above a pre-defined threshold (when 50% of tag-pairs are related) in Algorithm 1 line 1-5. The weight of the edge is set as the similarity value. Then, we assign a direction for each tag pair from the high centrality node to the low centrality node. As the graph is very likely not connected, we make the ROOT node point to every other node, with edge weight equal to the pre-defined threshold to make the graph weakly connected.

A directed rooted tree has a root vertex and exactly one directed path from root to any other vertex. A straightforward criteria is to find a tree that maximizes the edge weights. In essence, this follows a greedy strategy which was used previously by Heymann et al. (Heymann and Garcia-Molina 2006). They proposed to iteratively add nodes in a decreasing centrality order to a tree which maximizes similarity. From a graph point of view, we can apply Chu-Liu/Edmond's algorithm(Chu and Liu 1965) over the similarity graph. The core procedure is finding the edge incoming to node t of highest weight (with ties broken arbitrarily) for each t other than the ROOT. Since the edge order is predefined according to centrality, the graph is guaranteed to have no cycles and we can simplify the algorithm to forgo this cycle check.

Algorithm 1: Mincost Tree Formation

Input: Tag vectors
1 Calculate similarity between each pair of node (t_i, t_j)
2 Initialize directed weighted graph $\mathcal{G}$ with ROOT node
3 Add an edge (t_i, t_j) when $Sim(t_i, t_j) > threshold$
4 Assign direction for edges following centrality order
5 Add an edge between ROOT and each node with $weight = threshold$
6 **while** *n(edge)>n-1* **do**
7 **for** *each edge* (t_i, t_j) **do**
8 G' ← remove edge (t_i, t_j) from graph G
9 Find shortest path from t_i to t_j in G'
10 Calculate cost of deleting edge(t_i, t_j)
11 i=0
12 **while** *edge not removed* **do**
13 edge=increasing_cost_sequence[i]
14 **if** *the edge is not the last incoming edge to* t_j **then** remove edge from G and break;
15 i+=1
16 **return** $\mathcal{G}$

Generalized Cost Function Although different metrics are adopted for characterizing the relation between a pair of tags, they share the similar greedy strategy of minimizing the cost function

$$cost = \sum\nolimits_{e \in \mathcal{G}} W(e) - \sum\nolimits_{t \in \mathcal{G}} sim(t, t_p)$$

where $sim(t, t_p)$ represents the similarity between a tag t and its parent. Here we introduce a new minimum cost tree formation algorithm which builds upon the simplified Chu-Liu/Edmond's algorithm that generalizes the cost function. Concretely, the proposed folksonomy generation algorithm can be formalized as Algorithm 1.

After constructing a directed weighted graph $\mathcal{G}$, we next convert this graph to a tree $\mathcal{T}$ with minimum cost, where the cost here characterizes the structural change to $\mathcal{G}$. We define the cost for deleting the edge as $sim(t_i, t_j) \cdot d_{i,j}$, where $d_{i,j}$ represent the shortest path length from t_i to t_j in the graph which excludes the edge (t_i, t_j). After deleting an edge, the

two corresponding nodes are disconnected and we need to identify a new shortest route that connect these two nodes. The intuition is we want to maintain the coherence of the structure after deleting edges such that more similar tags tend to stay closer to each other. To do so, in each iteration, we calculate the cost of deleting each remaining edge in the graph $\mathcal{G}'$, and then find an edge with minimum cost which is not the last remaining edge pointing to the corresponding child node (so the node is not isolated). The algorithm stops when $n-1$ edges are left, with each node having exactly one parent. Thus our goal is to minimize the *structure conservation* cost function for converting $\mathcal{G}$ to a tree $\mathcal{T}$:

$$cost = \sum\nolimits_{e \in \mathcal{G}, e \notin \mathcal{T}} sim(t_i, t_j) \cdot d_{i,j}$$

However, this algorithm is computationally costly as whenever a new edge is deleted, it is required to recompute the shortest path between each pair of nodes. This is an O(E^2) shortest path calculation. Hence, we provide an approximation for the calculation shown in algorithm 2, where we only calculate the cost of deleting each edge in the original graph. According to the cost from low to high, we iteratively delete the edges until there is a unique parent for each node. Finally, the output is a location-sensitive folksonomy.

Algorithm 2: Approximation Algorithm

Input: Tag vectors
Constructing the Directed Weighed Graph $\mathcal{G}$ according to lines in Algorithm 1
for *each edge (t_i, t_j)* **do**
    G' $\leftarrow$ remove edge (t_i, t_j) from graph $\mathcal{G}$
    Find shortest path from t_i to t_j in G'
    Calculate cost of deleting edge(t_i, t_j)
while *n(edge)>n-1* **do**
    for *each edge in increasing cost* **do**
        if *the edge is not the last incoming edge to* t_j **then** Remove edge from $\mathcal{G}$;
return $\mathcal{G}$

Folksonomy-Informed Profiling

We turn in this section to apply the location-sensitive folksonomy for profile construction. We begin by finding candidate tags from folksonomy, and then embedding these candidates in a learning-to-rank framework for ordering the tags.

Finding Candidate Tags. Given a user's seen tag profile $P_s(u)$, we first leverage the location-sensitive folksonomy and select a set of candidate tags. To accomplish this, we locate each seen tag in $P_s(u)$ in the folksonomy and collect parent, children, and sibling tags of this seen tag as candidate tags. The hierarchical structure acts as a good filter and thus controls the number and quality of candidate tags. Then we order the candidate tags according to different strategies for prediction. The formal definition for this problem is given user u and a set of candidate tags $T_c(u) = \{t_1, ..., t_k\}$, we aim to find a scoring function to rank tags in $T_c(u)$ for u.

Features	Descriptions
$f(s)$	log scale overall frequency(freq) of the seed tag
$f(t)$	log scale overall freq of the tag
$f_u(s)$	log scale unique user freq of the seed tag
$f_u(t)$	log scale unique user freq of the tag
S_{sim}	similarity with seed tag
H_{sim}	highest similarity with existing tags
Hf_{sim}	H_{sim} weighted by freq
Sum_{sim}	sum of similarity with all seen tags
S^{w1}_{mv}	sum of similarity with existing seed tags weighted by $f(s)$
S^{w2}_{mv}	sum of similarity with existing seed tags weighted by $f_u(s)$
p_{cnt}	freq of the candidate tag being a parent
s_{cnt}	freq of the candidate tag being a sibling
c_{cnt}	freq of the candidate tag being a child

Table 1: Features for ranking candidate tags from the location-sensitive folksonomy.

Ranking Candidate Tags. We adopt a learning to rank approach for personalized candidate tag ranking. The advantage is that it automatically assigns optimum weight for each feature. We apply a pairwise learning algorithm RankSVM(Joachims 2002). Here we consider each user as a query and we assign each candidate tag an integer ranking score in the range of 3 to 1 according to its actual count in the user's unseen profile. RankSVM first generates a set of pairwise constraints and then transform the problem to a two-class classification problem according to those constrains and an SVM model is learned. Finally, in the ranking phase, rank scores are calculated according to the margin value. Note that we train the model with the training set and an L2 regularization term is added to prevent overfitting.

Here we introduce a set of features that we rely on to generate a preference order of the candidate tags for prediction in Table 1. A total of 13 features are used for training the model include features introduced above. Features can be grouped into three categories: user specific features, tag features, and folksonomy structure features. User specific features include $f_u(s)$, $f_u(t)$, H_{sim}, Sum_{sim}, Hf_{sim}, S^{w1}_{mv}, S^{w2}_{mv}. These features are retrieved from a user's seen profile, which represents characteristics of the user. Tag features includes $f(s)$, $f(t)$, and S_{sim}. These features only provide intrinsic properties of candidate tags. And folksonomy structure features include p_{cnt}, c_{cnt}, s_{cnt}, which are uniquely defined by the folksonomy to provide extra clues for making good predictions. The intuition here is that predicting a parent tag is more likely to be correct than a sibling or child, as parent tags are more general, having the largest overlap with the candidate. For example, inferring `football` to parent `sports` is more likely to be correct than to sibling `volleyball` and child `football player`.

Experiments

Data Preparation. We rely on a Twitter list dataset containing 15 million list relationships in which the geo-coordinates of the labelers and users are known (Cheng et al. 2014). In our experiments, the tags we included in the folksonomy

are extracted from each list name, and users in the list will be endowed with the tags in their profile. These tags contain multifaceted opinions of actual labelers, which means they can be complex and noisy. Hence, we apply text processing techniques such as case folding, stopword removal, and noun singularization. We also separate the string pattern like 'FoodDrink' into two words 'food' and 'drink'. We use language identification package (Lui and Baldwin 2012) to filter out non-English tags. To guarantee the informativeness and quality of the tags, we filter out infrequent tags with fewer than 5 labelers and 10 labelees. Twitter has a 25-character length limit for list names, and empirically we find nearly all list names do not exceed three words. We also include bigrams. Finally, the size of tagset is 10,489.

Profile Prediction Setup. For each of nine selected locations, a random sample of local users is held out. We construct a location-sensitive folksonomy given the location based on the rest of whole dataset. Following that we predict the user profiles for users in the hold-out data. For each user, the seen tag set $P_k(u)$ is a random 25% of his profile $P(u)$. Then we try to predict tags in the rest 75% unseen tags.[1] The result reported for every profiling experiment in this paper, including baselines, are based on four-fold cross validation and averaged over the nine locations.

Baselines. We consider two approaches based on collaborative filtering and Bayesian personalized ranking as baselines.

Collaborative Filtering-K Nearest Neighbor(CF-KNN). In CF-based profiling approach, we first identify the top-k local users that share the most similar tags with the target user. To maintain consistency with other approaches, we assume each user profile only contains 1/4 of the tags). Here, we apply cosine similarity to measure user similarity. Then, we aggregate the tags of the 50 nearest neighboring users weighted by their similarity and make predictions based on decreasing tag frequency in the collective neighbor profile.

Bayesian Personalized Ranking-Matrix Factorization(BPR-MF). We consider these tags as implicit feedback and our goal is identifying an optimal preference ranking of tags for each user. We thus experiment with two variations of state-of-the-art Bayesian personalized ranking criteria (Rendle et al. 2009). In the first setting, we train a unique model for each location by only considering its local users, denoted as "LBPR". We model a user i's affinity to tag j as $r_{ij} = p_i q_j + b_j$, where p_i and q_j represent latent factor of user and tag, respectively. b_j represents the overall preference of tag j. In the second setting, we train with whole dataset and explicitly model location-aware preferences, denoted as "LABPR". We define a user i's affinity to tag j as as $r_{ij} = p_i q_j + g_{l(i)j} + b_j$, where latent factor $g_{l(i)j}$ represents the regional popularity of tag j at the user i's home location (Lu and Caverlee 2015). For reproducibility, the number of negative samples, number of iterations, number of user and tag latent factors are set as 200, 80, 20 respectively. Regularization weights are set as 0.02.

[1] We only consider users with overall more than 10 tags.

Methods	P@1	P@5	AF@1	AF@5
Highest similarity	0.333†	0.285†	4.82†	4.24†
Freq. & similarity	0.507†	0.360†	14.6†	6.87†
Overall popularity	0.607†	0.501†	25.9	15.1†
$Sum_{similarity}$	0.651†	0.552†	25.3†	18.5
Learning to rank	0.763	0.677	26.5	19.2

Table 2: Comparing Tag Ranking Approaches. We observe that the LTR based approach results in the best precision, and also identifies the tags used most often (AF). '†' marks statistical significant difference with LTR according to paired t-test at 0.05.

Evaluation Metrics. The evaluation metrics we use are Precision@k (P@K) and Actual Frequency@k (AF@k). P@k measures how reliable predictions can be made. A high P@k value implies users have been labeled with the predicted tag, while high AF@k represents that users have been labeled many times with the predicted tags. Both measurements reported later are averaged over the test data. We consider the quality of prediction for the top-1 tag as well as top-5 tags. AF@k is defined as $\text{AF@}k = \sum_k f_u(t_k)/k$, where t_k is the kth predicted tag.

We now turn to the task of user profile construction based on location-sensitive folksonomies. We first compare the performance based on different ranking strategies, followed by profiling performance across different folksonomy generation approaches and local and general versions. Finally, we compare location-sensitive folksonomy informed profiling with the other baselines.

Comparing Ranking Strategies

Given the candidate tags identified from the folksonomy, our goal is to generate a personalized ranking over these tags so that the actual tags rank top. Here we compare the learning-to-rank (LTR) based approach with several baselines in Table 2, i). rank the candidate tags according to the decreasing order of similarity with the seen tag which subsumes the candidate tag. ii). rank with a hierarchical criteria, primarily according to the frequency of corresponding seen tag associated with the user and secondarily by decreasing order of similarity with the corresponding seen tag. iii). rank by overall tag popularity. iv). rank according to the aggregated similarity with the seen tag set.

We observe the LTR based approach outperforms all baselines in terms of precision, indicating high rank tags are more likely to be actual tags. Moreover, we find a similar trend in terms of AF@5, which represents the actual number of predicted tags that a user possesses. All these results imply the effectiveness of proposed features and feature weight scheme. Among the baselines, we find the "overall popularity" and "$Sum_{similarity}$" are relatively strong predictors.

Location Sensitive vs General Folksonomy

In Table 3, we compare the location-sensitive folksonomy versus the general folksonomy over all design choices for the application of profile construction. For each design choice, we acquire both location-sensitive and general version for

Methods	LS F.		G F.		LS F.	G F.
	P@1	P@5	P@1	P@5	AF	AF
MS-MC	0.754	0.663	0.653	0.571	19.2	15.2
SC-MC	0.763	0.677	0.656	0.571	19.2	15.4
MS-COS	0.751	0.656	0.662	0.521	16.4	11.9
SC-COS	0.756	0.663	0.662	0.531	18.1	12.8
MS-PMI	0.389	0.349	0.352	0.320	1.24	1.53
SC-PMI	0.402	0.362	0.363	0.332	1.21	1.62
MS-RBF	0.566	0.413	0.471	0.422	9.94	8.22
SC-RBF	0.581	0.430	0.465	0.410	10.1	12.0

Table 3: Comparing Location-Sensitive and General Folksonomies in Profile Tag Prediction. All location-sensitive versions are statistical significantly different with general versions according to paired t-test at 0.05.

each of the 9 locations mentioned in Table 3. The general folksonomy is constructed using the whole dataset excluding distance and location factors. The reported result is based on averaging 9 locations. We observe that overall, location-sensitive versions always beat its general counterpart in terms of precision@k and AF@5, regardless of design choice. The priority in terms of P@5 is around 0.1 and AF@5 is above by 20%. This result justifies the location-sensitive folksonomy since it better captures the local knowledge structures. It also demonstrates the effectiveness of how we model distance influence. Moreover, we find that the performance is consistent across locations.

We next compare the four similarity metrics used for constructing the tag similarity graph at the heart of location-sensitive folksonomy construction. Design choices with modified confidence (MC) perform best in terms of P@k, and AF@5, with folksonomy-informed version built on top of cosine similarity slightly lower, followed by RBF kernel and PMI. The extremely low performance of PMI indicates it may not be suitable for this scenario. After inspecting the folksonomy, we observe many of the hierarchical relations are incorrect or meaningless. PMI only considers co-occurrence of tags without taking the relative frequency difference into account. In our experiment design, as we don't set a minimum tag occurrence for each user to avoid sparsity, many tags only show up once on a user and it creates noise for an approach like PMI.

Last but not least, we evaluate the proposed *structure conservation* cost function (SC) against *maximum similarity* (MS) baseline. SC aims to construct a folksonomy that makes the smallest change to the similarity graph. We observe in Table 3 that applying SC leads to an incremental change in profile construction. For example, in the cosine case, we notice the P@5 and AF@5 are slightly better for SC. We notice about 9.5% of the relations are different among folksonomy generated using the two cost functions, meaning that the SC approach made some structure adjustments with some sacrifice in connecting most similar first strategy. Considering the limited difference in the two folksonomies, this increase in performance can be attributed to a better macro-structure.

Methods	P@1	P@5	AF@1	AF@5	tfidf@5
CF-KNN	0.656†	0.542†	25.6	18.0†	36.1†
LBPR	0.731†	0.650†	22.6†	12.3†	30.2†
LABPR	0.771	0.673	24.1†	16.2†	34.6†
LS Folk.	0.763	0.677	26.5	19.2	42.6

Table 4: Comparing Tag Prediction Approaches. '†' marks statistical significant difference with LS-Folk. approach according to paired t-test at 0.05.

Evaluate User Profiling

Finally, we compare location-sensitive folksonomy-informed user profiling with the CF-KNN and BPR-MF baselines. As we observe in Table 4, the proposed approach outperforms the CF-KNN and locally trained BPR-MF (LBPR) in both P@k and AF@k and exhibits similar performance compared with the location-aware BPR-MF (LABPR) approach. Even though LABPR does not consider frequency which is also important, the latent factors effectively capture user preferences over tags and location-based preference for tags. However, BPR-MF based approach is computationally costly as the dimension of user and tag increase. The CF-KNN approach is not robust in sparse condition, for example, when there are few similar users, the prediction made by CF-KNN could be very inaccurate.

We leverage average TFIDF score for top five predicted tags as a metric to reflect how important and informative the predicted tag is to a user in the actual tag collection. The score is averaged for users and locations. We notice the proposed LS-Folk yield the highest TFIDF@5, showing the capability of identifying uniquely important tag for the user. We notice that CF-KNN and BPR-MF based approach have a strong tendency to predict general high frequency tags. For CF-KNN, highly general tags are very likely to rank top in the sequence. For BPR-MF, the implicit feedback formulation neglects the difference in importance of the seen tags and has a tendency to predict tags that are seen on many users. These tags are often on a high abstraction level and thus provide only vague insight to a user. For example, if the predicted tag is `peep` which is short for "people", there is little new information contributed to the target user. In order to precisely profile a user, we expect to have concrete and specific tags, additionally, we wish to have a diverse tag space. In contrast, we observe that the location-sensitive folksonomy-informed approach performs much better in predicting diverse specific tags.

Conclusion

In this work, we explored the impact of spatial variation on the construction of location-sensitive user profile. Concretely, we proposed an location-sensitive folksonomy-informed framework toward the goal of improved user profiling. Through extensive experiments, we have demonstrated such location-sensitive folksonomy is more effective in identifying relevant tags, and learning to rank strategy is helpful for optimizing feature weights and leads to high quality user profile tags.

References

Ahmed, A.; Low, Y.; Aly, M.; Josifovski, V.; and Smola, A. J. 2011. Scalable distributed inference of dynamic user interests for behavioral targeting. In *SIGKDD*.

Backstrom, L.; Kleinberg, J.; Kumar, R.; and Novak, J. 2008. Spatial variation in search engine queries. In *WWW*.

Bhattacharya, P.; Ghosh, S.; Kulshrestha, J.; Mondal, M.; Zafar, M. B.; Ganguly, N.; and Gummadi, K. P. 2014. Deep twitter diving: Exploring topical groups in microblogs at scale. In *CSCW*.

Brodersen, A.; Scellato, S.; and Wattenhofer, M. 2012. Youtube around the world: geographic popularity of videos. In *WWW*.

Brooks, C. H., and Montanez, N. 2006. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW*.

Cheng, Z.; Caverlee, J.; Barthwal, H.; and Bachani, V. 2014. Who is the barbecue king of texas?: a geo-spatial approach to finding local experts on twitter. In *SIGIR*.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *CIKM*.

Chu, Y.-J., and Liu, T.-H. 1965. On shortest arborescence of a directed graph. *Scientia Sinica*.

Ghosh, S.; Sharma, N.; Benevenuto, F.; Ganguly, N.; and Gummadi, K. 2012. Cognos: crowdsourcing search for topic experts in microblogs. In *SIGIR*.

Heymann, P., and Garcia-Molina, H. 2006. Collaborative creation of communal hierarchical taxonomies in social tagging systems.

Heymann, P.; Ramage, D.; and Garcia-Molina, H. 2008. Social tag prediction. In *SIGIR*.

Hong, L.; Doumith, A. S.; and Davison, B. D. 2013. Co-factorization machines: modeling user interests and predicting individual decisions in twitter. In *WSDM*.

Hu, L.; Sun, A.; and Liu, Y. 2014. Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction. In *SIGIR*.

Ikeda, K.; Hattori, G.; Ono, C.; Asoh, H.; and Higashino, T. 2013. Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems*.

Jiang, M.; Cui, P.; Liu, R.; Yang, Q.; Wang, F.; Zhu, W.; and Yang, S. 2012. Social contextual recommendation. In *CIKM*.

Joachims, T. 2002. Optimizing search engines using clickthrough data. In *SIGKDD*.

Kaltenbrunner, A.; Scellato, S.; et al. 2012. Far from the eyes, close on the web: impact of geographic distance on online social interactions. In *Proceedings of the 2012 ACM workshop on online social networks*.

Krestel, R.; Fankhauser, P.; and Nejdl, W. 2009. Latent dirichlet allocation for tag recommendation. In *RecSys*.

Li, J.; Ritter, A.; and Hovy, E. H. 2014. Weakly supervised user profile extraction from twitter. In *ACL (1)*.

Li, R.; Wang, C.; and Chang, K. C.-C. 2014. User profiling in an ego network: co-profiling attributes and relationships. In *WWW*.

Liu, Q.; Chen, E.; Xiong, H.; Ding, C. H.; and Chen, J. 2012a. Enhancing collaborative filtering by user interest expansion via personalized ranking. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*.

Liu, X.; Song, Y.; Liu, S.; and Wang, H. 2012b. Automatic taxonomy construction from keywords. In *SIGKDD*.

Lu, H., and Caverlee, J. 2015. Exploiting geo-spatial preference for personalized expert recommendation. In *RecSys*.

Lui, M., and Baldwin, T. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*.

Majumder, A., and Shrivastava, N. 2013. Know your personalization: learning topic level personalization in online services. In *WWW*.

Plangprasopchok, A., and Lerman, K. 2009. Constructing folksonomies from user-specified relations on flickr. In *WWW*.

Qiu, F., and Cho, J. 2006. Automatic identification of user interest for personalized search. In *WWW*.

Rakesh, V.; Singh, D.; Vinzamuri, B.; and Reddy, C. K. 2014. Personalized recommendation of twitter lists using content and network information. In *ICWSM*.

Rego, A. S. C.; Marinho, L. B.; and Pires, C. E. S. 2015. A supervised learning approach to detect subsumption relations between tags in folksonomies. In *ACM Symposium on Applied Computing*.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*.

Ribeiro, I. S.; Santos, R. L.; Gonçalves, M. A.; and Laender, A. H. 2015. On tag recommendation for expertise profiling: A case study in the scientific domain. In *WSDM*.

Scellato, S.; Noulas, A.; Lambiotte, R.; and Mascolo, C. 2011. Socio-spatial properties of online location-based social networks.

Schmitz, P. 2006. Inducing ontology from flickr tags. In *Collaborative Web Tagging Workshop at WWW*.

Sigurbjornsson, B., and Van Zwol, R. 2008. Flickr tag recommendation based on collective knowledge. In *WWW*.

Song, Y.; Qiu, B.; and Farooq, U. 2011. Hierarchical tag visualization and application for tag recommendations. In *CIKM*.

Tuarob, S.; Pouchard, L. C.; and Giles, C. L. 2013. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital Libraries*.

Verma, C.; Mahadevan, V.; Rasiwasia, N.; Aggarwal, G.; Kant, R.; Jaimes, A.; and Dey, S. 2015. Construction and evaluation of ontological tag trees. *Expert Systems with Applications*.

Wang, S.; Tang, J.; Wang, Y.; and Liu, H. 2015. Exploring implicit hierarchical structures for recommender systems. In *IJCAI*.

Weng, J.; Lim, E.-P.; Jiang, J.; and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*.

Xia, X.; Lo, D.; Wang, X.; and Zhou, B. 2013. Tag recommendation in software information sites. In *MSR*.

Yin, H.; Cui, B.; Chen, L.; Hu, Z.; and Huang, Z. 2014. A temporal context-aware model for user behavior modeling in social media systems. In *SIGMOD*.

Zhang, H.; Korayem, M.; You, E.; and Crandall, D. J. 2012. Beyond co-occurrence: discovering and visualizing tag relationships from geo-spatial and temporal similarities. In *WSDM*.

Zhao, Z.; Cheng, Z.; Hong, L.; and Chi, E. H. 2015. Improving user topic interest profiles by behavior factorization. In *WWW*.

Zhong, E.; Liu, N.; Shi, Y.; and Rajan, S. 2015. Building discriminative user profiles for large-scale content recommendation. In *SIGKDD*.

Zhu, X.; Ming, Z.-Y.; Hao, Y.; and Zhu, X. 2015. Tackling data sparseness in recommendation using social media based topic hierarchy modeling. In *IJCAI*.