Towards Diversified Local Users Identification Using Location Based Social Networks

Chao Huang, Dong Wang, Shenglong Zhu
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556
{chuang7,dwang5,szhu3}@nd.edu

***Abstract*—Identifying a set of diversified users who are local residents in a city is an important task for a wide spectrum of applications such as target ads of local business, surveys and interviews, and personalized recommendations. While many previous studies have investigated the problem of identifying the local users in a given area using online social network information (e.g., geotagged posts), few methods have been developed to solve the diversified user identification problem. In this paper, we propose a new analytical framework, *Diversified Local Users Finder (DLUF)*, to accurately identify a set of diversified local users using a principled approach. In particular, the DLUF scheme first defines a new distance metric that measures the diversity between local users from physical dimension. The DLUF scheme then provides an optimal solution to find the set of local users with maximum diversity. The performance of DLUF scheme is compared to several representative baselines using two real world datasets obtained from Foursquare application. We observe that the DLUF scheme accurately identifies the local users with a great diversity and significantly outperforms the compared baselines.**

***Index Terms*—Diversified Local Users, Location Based Social Networks, Foursquare**

I. Introduction

Location-Based Social Network (LBSN) services have become a popular paradigm for people to share their "check-in" traces (i.e., a sequence of GPS coordinates at places the users visited) online [7]. Examples of LBSNs include Foursquare, Google Places, Gowalla, Yelp, and Wechat. In this paper, we study the problem of accurately identifying a set of *diversified* users in a city from the check-in traces they voluntarily share on location based social networks. We define the location-based diversity of users as the geographic difference between users home locations, which can be inferred from the data of LBSNs.

Identifying the set of diversified local users in a city is important for many information services and applications [6], [18], [21], [20]. For example, the local business can leverage the diversified set of users to implement target ads to maximize its opportunity to reach potential customers [1]. Alternatively, the local government and communities can use the diversified set of users as good candidates for their surveys and interviews that prefer independent responses and population diversity [5].

In this paper, we develop a new analytical framework, *Diversified Local Users Finder (DLUF)* scheme to solve the diversified local user identification problem. In particular, we first estimate the users' home locations by mining their online check-in traces using an *unsupervised* estimation framework. Then, we introduce a new distance metric to compute the diversity scores between users from physical dimension (i.e., geographical distance between their home locations). Finally, we formulate the diversified local user identification problem as a constraint optimization problem and derive the solution to find the set of local users with a maximum diversity between them. In evaluation, we study the performance of DLUF scheme using two real world datasets obtained from Foursquare. We observe that the DLUF scheme accurately identifies the local users with a great diversity and significantly outperforms the compared baselines.

In summary, the contributions of this paper are as follows:

- We are among the first to study the problem of identifying a set of local diversified users based on their check-in traces from LBSNs.
- We develop a principled framework (i.e., DLUF scheme) that allows us to accurately identify a set of local users with maximum diversity (Section IV)
- We perform extensive experiments to study the performance of our DLUF scheme and compare it with several baselines using two real world datasets obtained from Foursquare. The results demonstrate the effectiveness of the DLUF scheme. (Section V)

The remainder of this paper is organized as follows. We first review related work in Section II. Section III formally define the problem of this work. Then, we present our solution in Section IV. In Section V, we present experimental results. We conclude this work in Section VI.

II. Related Work

User Profiling. There exists a good amount of works that focus on the problem of user profiling [10], [4], [12], [15],

ASONAM'17, July 31 - August 03, 2017, Sydney, Australia

ACM ISBN 978-1-4503-4993-2/17/07?/$15.00
http://dx.doi.org/10.1145/3110025.3110159

[17], [16]. In particular, Ikeda et al. developed a hybrid estimation scheme to profile users on Twitter by analyzing both their tweets and social relationships [10]. Geng et al. studied the problem of content-based user profiling in social curation services by exploring the content-based user's preference and social relationship [4]. Li et al. inferred users' attributes by considering their social connections based on ego networks and the dependency between user's attributes [12]. Mislove et al. inferred user's missing profile attributes on Facebook by exploring the attributes of user's social friends based on the online community structures [15]. In this work, we study a new problem of identifying the diversified set of local users by estimating their home locations and maximizing the diversity between the identified users.

Geo-locating User. User geo-locating in a city is a critical challenge in social network applications and previous works have made a significant progress to address this problem [14], [2], [7]. For example, McGee et al. developed a network-based scheme to infer user's locations by leveraging the strength of social ties between users on social media platform [14]. Backstrom et al. explored the social and geographical dependencies between users to estimate their locations [2]. Huang et al.developed a spatial-temporal-social aware framework to identify the localness of users in a city based on their online check-in traces [7]. They futher extended their framework to infer the family relationship between users in a city based on the inferred home locations [8]. In this paper, we solve a different diversified user identification problem where the goal is to find a set of users with maximum diversity defined on the physical dimension.

III. Problem Formulation

In this section, we formally define the problem of identifying a set of diversified users in a city using data from LBSNs. We introduce the basic terminologies to be used in this problem formulation as follows.

1) Basic Terminology: We first define the set of selected local users whose diversity is the maximum as the *optimal diversity set*. The number of users in the optimal diversity set is denoted as k[1]. We denote the number of venue categories (e.g., restaurant, entertainment) as C which is indexed by t. In addition, we define the inputs to our scheme as follows.

Definition 1. **Physical Distance Matrix HD.** We define Physical Distance Matrix $HD_{Y\times Y}$ to represent the geographical distance between home locations of each pair of users in U. Specifically, in HD, $HD_{y,y'}$ represents the physical distance between the home location coordinates of user U_y and $U_{y'}$.

Definition 2. **Diversity Score Matrix DS.** We define Diversity Score Matrix $DS_{Y\times Y}$ to represent the diversity score between each pair of users in U. Specifically, in DS, $DS_{y,y'}$ is the diversity score between user U_y and $U_{y'}$: $DS_{y,y'} = \frac{HD_{y.y'}}{max(HD)}$. In this paper, we assume the distance score is symmetric (i.e., $DS_{y,y'} = DS_{y',y}$).

[1]The number of users in the set is often decided by the applications based on various factors (e.g., targeted influence scope, budget limit, etc.)

2) Problem Statement: Given a location coordinates l (i.e., lat and lon), a distance radius r, and a category of venues b (e.g., a local business may only be interested in users who live within X miles from its location and visited the same category of venues as its own.), our objective is to find the optimal set of k local users D^* who visited venues of an interested category b and live at a place within the distance r from l, and whose *diversity score* among k users is the maximum. Mathematically, this can be expressed as follows:

$$\begin{aligned} max \ & \sum_{y=1}^{Y}\sum_{y'\neq y} DS_{y,y'}\cdot\lambda_y\cdot\lambda_{y'} \\ s.t \ & \lambda_y \in \{0,1\} \\ & \Sigma_{y=1}^{Y}\lambda_y = k \\ & b \subset V_{U_y} \\ & \text{distance}(lat, lon, \varphi_u, \delta_u) \leq r \end{aligned} \quad (1)$$

where $\lambda_y = 1$ (or 0) indicates that user U_y is selected (or not) and V_{U_y} represents the set of venue categories user U_y has visited. $DS_{y,y'}$ represents the diversity score between user U_y and $U_{y'}$.

IV. The Diversified Local Users Identification Framework

In this section, we present the *Diversified Local Users Finder (DLUF)* scheme to find the optimal diversity set by maximizing the average diversity score between the selected users.

We formulate the problem of finding the optimal diversity set of users (whose home location diversity is maximized) as an optimization problem and solve it using the integer programming approach. Based on the outputs of home location inference scheme [9], we firstly select the users who meet the category and distance constraints defined in Equation (1). In particular, we select the users who visited venues of category b in their data traces and whose inferred home location from home location estimation approach are within radius r from the center l given by the application. We denote the initial candidate set of users selected using the above rule as D. We then compute the Diversity Score matrix DS for each pair of users in D using the distance metric defined in *Definition 1*. Based on DS matrix, we construct an undirected graph $G_{ds} = (V_{ds}, E_{ds}, W_{ds})$ where V_{ds} represents the initial candidates in D, E_{ds} represents their diversity relationship and W_{ds} represents the *diversity score* in DS. We let e_j represent the edges in E_{ds} and w_j represent the weight of e_j.

Given an integer k (i.e., the size of optimal diversity set), the objective is to find a set of vertices V^* of size k such that the sum of the edge weights in the subgraph represented by V^* is maximized. This problem is NP-hard.

The above problem can be formulated as an Integer Linear Programming (ILP) problem [3]. For each vertex $v_i \in V_{ds}$, we define a variable η_i such that $\eta_i = 1$ if and only if v_i is chosen in the solution. For each edge $e_j \in E_{ds}$, we define a variable ϕ_j such that $\phi_j = 1$ if and only if e_j is in the

induced subgraph of the chosen vertices. We further define w_j to represent the weight of edge e_j. Since the objective is to maximize the sum of edge weights in the induced subgraph, the objective function can be written as $\sum_{j=1}^{m} w_j \cdot \phi_j$, where $m = |E_{ds}|$. We let $n = |V_{ds}|$. Formally, the problem can be formulated as follows:

$$\begin{aligned} \max \quad & \sum_{j=1}^{m} w_j \cdot \phi_j \\ \text{s.t.} \quad & \sum_{i=1}^{|n|} \eta_i = k \\ & \phi_j \leq (\eta_p + \eta_q)/2, \quad \text{for all } e_j = (v_p, v_q) \in E_{ds} \\ & \eta_i \in \{0,1\}, \qquad i = 1, \ldots, n \\ & \phi_j \in \{0,1\}, \qquad j = 1, \ldots, m \end{aligned} \tag{2}$$

We solve this problem by using Integer Linear Programming (ILP) [13] (i.e., $ILP\text{-}Func()$).

V. Evaluation

In this section, we study the performance of the *DLUF (Diversified Local Users Finder)* scheme by carrying out experiments on real world datasets collected from Foursquare, a location-based social network service. We compare the performance of the *DLUF* scheme with the state-of-the-art baselines and show that our scheme achieves non-trivial performance gain on maximizing the diversity of selected users.

A. Data Traces and Evaluation Metrics

1) Data Trace Statistics: In our evaluation, we use two data traces obtained from the Foursquare platform. In Foursquare, users are able to post their locations by checking-in at venues they visit. Essentially, a user's check-in point can be represented as: (user ID, venue ID, timestamp). In this paper, we targeted at users' data traces from two cities in U.S: Washington D.C. and Chicago. Table I presents the statistics of the two data traces. One should note that we only use such ground truth information on user's home locations to evaluate the performance of DLUF and do not use it as the input to our scheme. Instead, we use the *inferred* home locations of users estimated from DLUF scheme to compute the Diversity Score Matrix DS for our optimal diversified local user identification scheme.

Table I
Statistics of Two Data Traces

Data Trace	Washington D.C	Chicago
Number of Users	17,231	31,965
Number of Venues	1,932	2,529
Number of Check-ins	25,722	48,605

2) Evaluation Metric: We define *Average Diversity Score* between k users in a set D as:

$$Ave_k = \frac{(\Sigma_{U_x \in S^*} \Sigma_{U_{x'} \in S^*} DS_{x,x'})}{C_k^2} \tag{3}$$

where $C_k^2 = \frac{k \cdot (k-1)}{2}$. The higher the score is, the more diversified users are in the selected set.

B. Evaluation Results

Table II
Average Diversity Score of Selected k Users on Washington D.C. Data Trace ($k = 10$)

Algorithm	Radius (miles) $r = 4$	$r = 6$	$r = 8$
DLUF	**0.174**	**0.229**	**0.250**
CGA	0.102	0.140	0.164
TkLUS	0.081	0.102	0.158
Raw	0.039	0.045	0.086

Table III
Average Diversity Score of Selected k Users on Washington D.C. Data Trace ($k = 15$)

Algorithm	Radius (miles) $r = 4$	$r = 6$	$r = 8$
DLUF	**0.150**	**0.201**	**0.207**
CGA	0.081	0.132	0.151
TkLUS	0.060	0.093	0.104
Raw	0.043	0.053	0.066

Table IV
Average Diversity Score of Selected k Users on Chicago Data Trace ($k = 10$)

Algorithm	Radius (miles) $r = 4$	$r = 6$	$r = 8$
DLUF	**0.147**	**0.199**	**0.206**
CGA	0.109	0.121	0.132
TkLUS	0.088	0.097	0.109
Raw	0.068	0.072	0.077

Table V
Average Diversity Score of Selected k Users on Chicago Data Trace ($k = 15$)

Algorithm	Radius (miles) $r = 4$	$r = 6$	$r = 8$
DLUF	**0.131**	**0.184**	**0.208**
CGA	0.087	0.115	0.156
TkLUS	0.078	0.081	0.102
Raw	0.047	0.062	0.073

1) Evaluation of Diversified Local Users Finding: In this subsection, we evaluate the performance of the proposed *DLUF* scheme to identify the optimal diversity set of local users and compare it to three user selection techniques that include:

- *TkLUS*: it proposes a method that integrates user's location information for user selection [11].
- *CGA*: it selects users by exploring the structure property of social networks between users [19].
- *Raw*: it randomly select k users from the candidates.

In our evaluation, we choose the food category as our interested categories. In the experiment, we selected the center l of our interested area as the center of the city and varied both the radius r and the number of selected users (i.e., the size of the optimal diversity set) k. The results of Washington D.C. data trace are presented in Table II and Table III, respectively. We observe that our DLUF scheme outperforms all compared schemes by selecting k users with larger average diversity scores (i.e., finding a set of more diversified users). We also observe that the performance gain is stable for different values of k and r, which indicates a robust performance of the DLUF scheme on the key parameters of the problem. Additionally, Table IV and Table V present the evaluation results on Chicago trace. We observed that the DULF scheme continues to outperform the compared baselines in terms of finding more diversified users.

VI. Conclusion

This paper proposes a principled framework to identify a set of diversified local users by mining the publicly available data from LBSN. In particular, we develop the *Diversified Local User Finder (DLUF)* scheme that can accurately identify the optimal set of diversified users by maximizing their diversity score computed from the physical dimension. The performance of DLUF scheme is evaluated using two real world data traces obtained from Foursquare application. The evaluation results demonstrated that the DLUF scheme significantly outperforms current solutions by finding a set of more diversified users in the given area of interests. The diversified set of users can be used in many information services such as target ads of local business, surveys and interviews, and personalized recommendation systems.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. CBET-1637251, CNS-1566465 and IIS-1447795 and Army Research Office under Grant W911NF-16-1-0388. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] A. Ahmed, Y. Low, M. Aly, V. Josifovski, and A. J. Smola. Scalable distributed inference of dynamic user interests for behavioral targeting. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 114–122. ACM, 2011.

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

[3] R. S. Garfinkel and G. L. Nemhauser. *Integer programming*, volume 4. Wiley New York, 1972.

[4] X. Geng, H. Zhang, Z. Song, Y. Yang, H. Luan, and T.-S. Chua. One of a kind: User profiling by social curation. In *Proceedings of International Conference on Multimedia (MM)*, pages 567–576. ACM, 2014.

[5] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, pages 779–782, 2008.

[6] C. Huang and D. Wang. Spatial-temporal aware truth finding in big data social sensing applications. In *Trustcom/BigDataSE/ISPA*, volume 2, pages 72–79. IEEE, 2015.

[7] C. Huang and D. Wang. Exploiting spatial-temporal-social constraints for localness inference using online social media. In *Proceedings of International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 287–294. IEEE, 2016.

[8] C. Huang, D. Wang, S. Zhu, and B. Mann. Toward local family relationship discovery in location-based social network. *Social Network Analysis and Mining (SNAM)*, 2017.

[9] C. Huang, D. Wang, S. Zhu, and D. Y. Zhang. Towards unsupervised home location inference from online social media. In *Proceedings of International Conference on Big Data (Big Data)*, pages 676–685. IEEE, 2016.

[10] K. Ikeda, G. Hattori, C. Ono, H. Asoh, and T. Higashino. Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems (KBS)*, 51:35–47, 2013.

[11] J. Jiang, H. Lu, B. Yang, and B. Cui. Finding top-k local users in geo-tagged social media data. In *Proceedings of International Conference on Data Engineering (ICDE)*, pages 267–278. IEEE, 2015.

[12] R. Li, C. Wang, and K. C.-C. Chang. User profiling in an ego network: co-profiling attributes and relationships. In *Proceedings of International Conference on World Wide Web (WWW)*, pages 819–830. ACM, 2014.

[13] A. Makhorin. Glpk (gnu linear programming kit), 2008.

[14] J. McGee, J. Caverlee, and Z. Cheng. Location prediction in social media based on tie strength. In *Proceedings of International Conference on Information and Knowledge Management (CIKM)*, pages 459–468. ACM, 2013.

[15] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of International Conference on Web Search and Data Mining (WSDM)*, pages 251–260. ACM, 2010.

[16] D. Wang, M. T. Al Amin, T. Abdelzaher, D. Roth, C. R. Voss, L. M. Kaplan, S. Tratz, J. Laoudi, and D. Briesch. Provenance-assisted classification in social networks. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):624–637, 2014.

[17] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti, et al. Using humans as sensors: an estimation-theoretic perspective. In *Information Processing in Sensor Networks, IPSN-14 Proceedings of the 13th International Symposium on*, pages 35–46. IEEE, 2014.

[18] D. Wang and C. Huang. Confidence-aware truth estimation in social sensing applications. In *Proceedings of International Conference on Sensing, Communication, and Networking (SECON)*, pages 336–344. IEEE, 2015.

[19] Y. Wang, C. Cong, G. Song, and K. Xie. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1039–1048. ACM, 2010.

[20] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 325–334. ACM, 2011.

[21] D. Y. Zhang, R. Han, D. Wang, and C. Huang. On robust truth discovery in sparse social media sensing. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 1076–1081. IEEE, 2016.