Topic Modeling-based Semantic Annotation of Place using Personal Behavior and Environmental Features

Yohan Chon, Yungeun Kim, Hyojeong Shin, Hojung Cha
Department of Computer Science
Yonsei University, Seoul, Korea
{yohan,ygkim,hjshin,hjcha}@cs.yonsei.ac.kr

ABSTRACT

Understanding the semantics of the places is critical to improve emerging mobile services. In this paper, we present topic modeling-based place characterizing method to solve the problem of *Dedicated Task 1* in *Nokia Mobile Data Challenge*: semantic annotation of place. We applies the principles of topic modeling to leverage the context of smartphone users to infer place categories. In the proposed approach, topics are considered as the semantic category of places (e.g., home, workplace, restaurant), places are modeled as documents, and the personal behavior pattern (e.g., mobility pattern, calling or messaging, phone-silent setting) and environmental features (e.g., number of radio beacons) are discretized into terms that populate each document. Our results show that the proposed method can classify places into 10 categories with an overall accuracy of 68%.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; J.4 [**Computer Applications**]: Social and Behavior Sciences.

General Terms

Algorithms, Design, Experimentation, Human Factors

Keywords

Semantic Location, Location-based Services

1. INTRODUCTION

Understanding the semantics of the places is critical to improve emerging mobile services. While extensive attempts have been made to infer type-of-places in everyday life, previous work has a lack of inferring various type-of-places. Ye et al [11] used check-in behavior of users at social network (e.g., Whrrl) for the semantic annotation of places, but they focused on only three categories (i.e., restaurant, shopping, and nightlife). Works in [9] used image data to infer semantic of places in home space. They found that objects in places are correlated with the semantic labeling of placs, but this work also have characterized only four categories (i.e., bathroom, bedroom, kitchen, office). Chon et al [2] used image and audio data captured by mobile phones to characterize places into 7 categories. Isaacman et al [5] employed call detail records of smartphone users to identify important places such as home and workplace. Consequently, semantic annotation of places with various type-of-places remains to be solved in the mobile computing area.

Table 1: Description of data set

Place Category	Train Set		Test Set	
	Place	Visit	Place	Visit
Unknown	6014	24976	3043	26724
My home	84	14168	-	-
Others' home	46	1869	-	-
My workplace/school	102	9721	-	-
Transportation	23	110	-	-
Others' workplace/school	9	167	-	-
Outdoor sports	25	343	-	-
Indoor sports	14	468	-	-
Restaurant or bar	11	303	-	-
Shopping place	17	295	-	-
Holiday or vacation spot	5	6	-	-

The *Nokia Mobile Data Challenge* announes the contest for solving the problem of semantic annotation of places. Nokia Research Center Lausanne and its academic Swiss partners have collected smartphone data from almost 200 participants in the course of 1+ year [6]. Table 1 describes the collected dataset. *Train Dataset* includes 6350 places visited by 80 users and 336 places are categorized into 10 place categories. *Test Dataset* contains 3043 places visited by 34 users and none-places are labeled. The problem is the classification into 10 classes. The problem is very challenging and practical compared with previous works that focused on 2 to 4 categories [11, 9, 5].

In this paper, we present a topic modeling-based place characterizing method to solve the problem of *Dedicated Task 1* in *Nokia Mobile Data Challenge*: semantic annotation of place. We applies the principles of topic modeling to leverage the context of smartphone users to infer place categories. In proposed approach, topics are considered as the semantic category of places (e.g., home, workplace, and restaurant) and places are modeled as documents. We analyzed various contextual information collected by smartphone users to explore meaningful features related to type-of-places. We discretized personal behavior patterns (e.g., mobility pattern, calling or messaging, phone-silent setting) and environmental features (e.g., number of bluetooth devices, number of WiFi APs, phone-charging opportunities) into terms that populate each document. We explored the meaningful features related to type-of-places and evaluated proposed method with supervised- and semi-supervised-approach. The results showed that the proposed

This material was prepared for the Mobile Data Challenge 2012 (by Nokia) Workshop; June 18-19, 2012; Newcastle, UK. The copyright belongs to the authors of this paper.

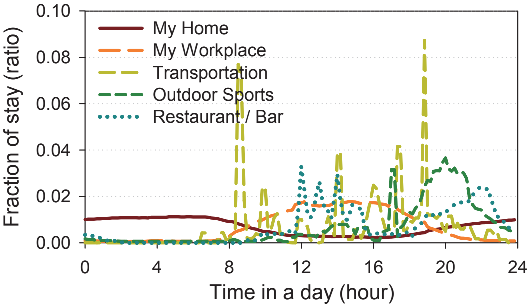

Figure 1: Residence-time distribution at different type-of-places.

method can classify places into 10 categories with an overall accuracy of 68%.

2. SEMANTIC ANNOTATION OF PLACES

We applies the principles of topic modeling to leverage the various features for inferring place categories. We first describes the features considered as hints that infer place categories. We then present the topic modeling method with two phase classification.

2.1 Personal Behavior Patterns

The basic assumption in the use of personal behavior patterns is that people tend to show similar behavior at places of same categories. For example, people tend to visit restaurant around meal times or they set the phone into silence-mode in the meeting room. To differentiate type-of-places, the system should investigate meaningful features at specific type-of-place across all users.

Mobility Pattern. The underlying assumption in the use of user trajectories is that the time of day when people visit certain place categories has a consistent pattern. Intuitive examples include a person spending meal times at food-related places or people are often found at their workplace on weekdays between 9 to 5. Consider a user's mobility history $H = (l_1, t_1^a, s_1), \cdots, (l_n, t_n^a, s_n)$, in which t_i^a is the arrival time and s_i is the stay-duration at location l_i. From H, we extract the residence time history $(t_1^a, s_1), \cdots, (t_m^a, s_m)$ at a specific location l^k. Then, the residence-time distribution R at location l^k is a form of a discrete histogram distribution from the set of residence time $\mathbb{R}^k = \{(t_x^a, s_x) \mid l_i = l^k\}$. Here, each histogram bin represents a certain period (e.g., 10 minutes) during a single day. We build two sets of residence-time distributions, one for the weekend and one for weekdays as suggested in [10]. Figure 1 presents encoded residence-time distributions on a weekday. Intuitively, participants spent their nights at home, and they spent work hours at the workplace. The peaks of R at restaurant/bar are observed around meal times (12pm) and nighttime (10pm). The result indicates that mobility patterns at different type-of-place are meaningful features for differentiating type-of-places.

We found that visit frequency and stay duration at places are reasonable features to differentiate type-of-places. Intuitively, people tend to spent a majority of their time and frequently visits a few major-places (e.g., home and workplace) [3]. In other words, visit frequency at home or workplace is significantly larger than frequency at shopping or restaurant places. We defined visit frequency as the fraction of visited days to total collection period: $\frac{\sharp \text{ of visit days}}{\sharp \text{ of total days}}$.

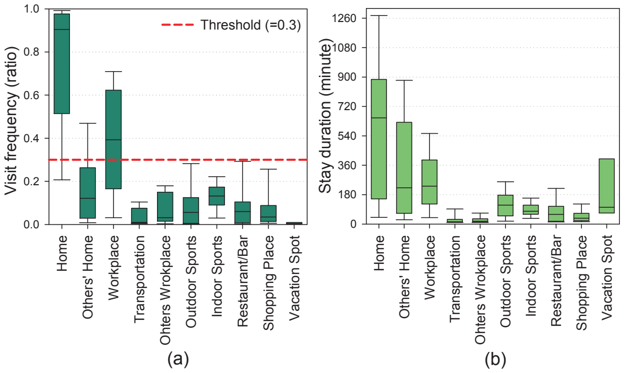

Figure 2: Boxplot of (a) visit frequency and (b) stay duration. Box indicates 25%, 50%, and 75% of data, and whisker indicates 10% and 90%.

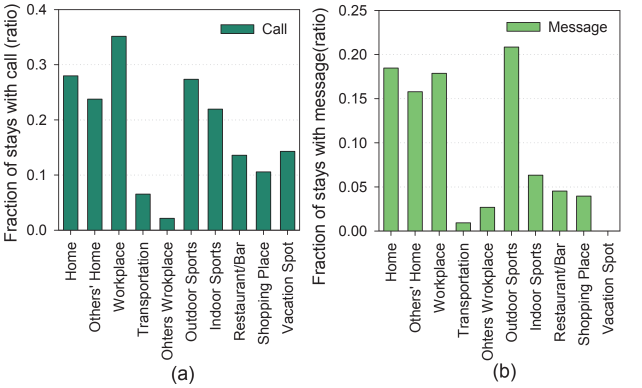

Figure 3: Ratio of stay with (a) call and (b) message.

Figure 2 presents the distribution of visit frequency and stay duration at different type-of-places in dataset. The result indicates that the visit frequency can be used to extract major places (i.e., home or workplace) from all places, as shown in Figure 2(a). Additionally, participants tend to stay for a longer duration at home or workplace than other places. We used visit frequency to choose major places in the first phase (see Section 2.3), and discretized stay duration into 9 bins (i.e., 15, 30, 60, 90, 120, 240, 360, 600, and 600+ minutes).

Calling and Messaging. We estimated the ratio of calling and messaging behavior at each place, defined as $\frac{\sharp \text{ of stays with calling or messaging}}{\sharp \text{ of total stays at a place}}$. Here, 0.3 indicates that participants used calling/messaging 3 times out of 10 times during visits at a specific place. We found that certain type-of-places (e.g., others' workplace or restaurant) may not elicit calling and messaging behavior, as shown in Figure 3. The main reason is that people tend to not interact with their smratphones at those places since these places are directly connected to specific activities with other people (e.g., having meeting with co-workers or meal with friends). We discretized the ratio of calling/message as terms in the document.

Phone Setting. We expect that people tend to set phone into silent-mode at specific type-of-places. The dataset indicates that participants used silent-mode at about 40% of stays at workplace or indoor sports, as shown in Figure 4(a). Similar to calling/messaging behavior, we estimated the ratio of silent-mode at each place, and generated it into terms for topic modeling.

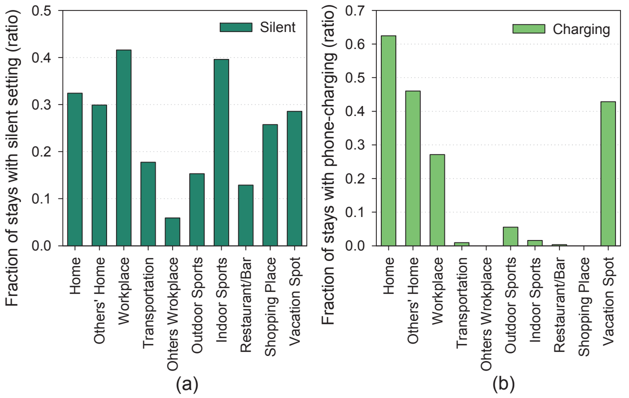

Figure 4: Ratio of stay with (a) silent setting and (b) phone-charging.

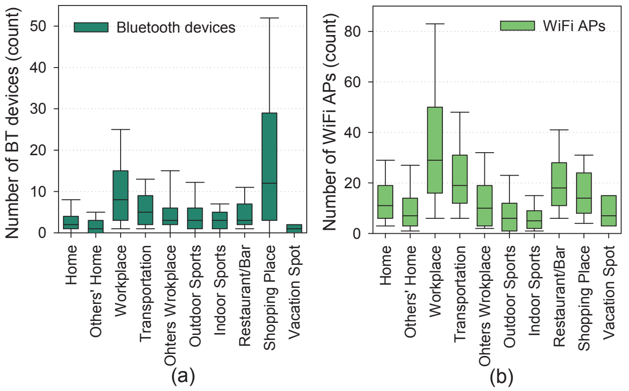

Figure 5: Boxplot of (a) number of bluetooth devices and (b) number of WiFi APs.

2.2 Environmental Features

We explored environmental features of places to estimate the discrimination power of features. The basic assumption is that places in the same category have similar infrastructures or population densities. Intuitive example is that home places may be equipped with a battery charger or shopping places are crowded with people.

Charging Opportunity. We expected that private places (e.g., home or workplace) offers charging opportunity for smartphone, but people cannot charge their phones at restaurants or shopping places. Figure 4(b) clearly shows this tendency. Intuitively, participants charged their smartphone at home places for about 60% of stays. Participants mainly charged their smartphones at home, workplace, or vacation spot. We estimated the ratio of charging opportunity at each stay behavior of places, and transformed it into terms.

Density of People. We used the number of Bluetooth (BT) devices scanned at places to infer the density of people. The underlying assumption is that the density of public space (e.g., shopping or workplace) is higher than those of private space. Figure 5(a) presents the distribution of the number of BT devices at the different type-of-places. The shopping place shows the highest number of BT devices because of crowd-situation while home space shows the smallest number of BT devices. We descritized the number of BT devices into 6 bins (i.e., 2, 5, 10, 15, 20, and 20+ devices).

Radio Beacons. The infrastructures at places indicate the characteristics of type-of-places. With the increasing usage of wireless network, most workplace or service-places (e.g., restaurant, bar, or shopping place) set up WiFi APs for work or service purposes. Figure 5(b) presents the number of WiFi APs at various type-of-places in the dataset. The places shared with many people (e.g., workplace, transporation, restaurant/bar, shopping places) contains a larger number of WiFi APs than private places (e.g., home) or places for sports. We used 7 bins (i.e., 5, 10, 15, 20, 30, 40, and 40+ APs) as descritized terms for radio beacons.

2.3 Place Categorization

We chose topic modeling method as a classifier for leveraging multi-modal features. Topic modeling is a widely used statistical model for discovering the abstract topics that occur in a collection of documents [4]. We employed the Labeled Latent Dirichlet Allocation (L-LDA) model [7]. L-LDA is an extension of traditional LDA [1], allowing topic models to be trained with labeled documents and multiple labels: the place can have more than one type-of-place. These characterisitics are appropriate for mobile sensing scenario since two users may consider the same place as different type-of-places. The proposed method estimates the topic-specific distribution from the labeled places, and infer the type-of-place in unlabeled places.

We now briefly overview the training process of the L-LDA model to extract topics (place categories) from collection of documents (places). Topics are learned from the co-occurance terms in places from the same category. Let each document d be represented by a tuple consisting of a list of word indices $w^{(d)} = (w_1, \cdots, w_{N_d})$ and a list of binary topic presence/absence indicators $\Lambda^{(d)} = (p_1, \cdots, p_K)$ where each $w_i \in \{1, \cdots, V\}$ and each $p_k \in \{0, 1\}$. Here N_d is the document length, V is the size of vocabulary that includes all classifier terms and user trajectory terms, and K is the total number of unique labels in the corpus. The model generates multinomial topic distributions over vocabulary $\beta_k = (\beta_{k,1}, \cdots, \beta_{k,V})^T \sim \mathrm{Dir}(\cdot \mid \eta)$ for each topic k, from a Dirichlet prior η. The L-LDA model then draws a multinomial mixture distribution $\theta^{(d)}$ over the topics that correspond to their labels $\Lambda^{(d)}$. For any document, the final topic distribution $\theta^{(d)}$ will correspond to the relevance of the topic within the document. In other words, $\theta^{(d)}$ will indicate the strength of a place category.

We used term frequency-inverse document frequency (tf-idf) [8] to determine terms that are rare across all places. In case of personal behavior patterns, we applied tf-idf to the corpus of each user to preserve the personal characteristics. We estimate unique behavioral patterns at specific places within individual data. In case of environmental features, we consider the documents of all users as one corpus to determine unique terms in global spaces.

Two Phases Classification. We designed two phase classification for differentiating major places and minor places. We defined major places as the places people spend most of their times in daily life such as home or workplace. The main reason of two phases classification is (1) to reduce the number of categories within a subset of data and (2) to separate long-length documents (major places) with short-length documents (minor places). We used the visit frequency to divide visited places into major and minor places. Then, the classification problem at first phases considers major places with three categories: home, workplace, and others. In second phase, we filtered out places recognized as home or work-

Table 2: Summary of method.

Topic Model	Train Set	1st Phase - 3 categories	2nd Phase - 10 categories
L-LDA	Labeled Data	Mobility+Env.*	Mobility+Env.
			Mobility+Env.+Prsnl.*
LDA	Labeled Data+ Unlabeled Data	Mobility+Env.	Mobility+Env.
			Mobility+Env.+Prsnl.

* Env: Environmental features, Prsnl: Calling/messaing/phone setting.

Table 3: Confusion matrix at first phase.

Result Ground truth	My home	My workplace	Others
My home (71)*	0.96	0.00	0.04
My workplace (59)	0.00	0.93	0.07
Others (9)	0.11	0.11	0.78

* ( ) indicates a number of places

place at first phase, and classify the remaining places into 10 place categories.

3. EVALUATION

We submitted five solutions with different settings as shown in Table 2. Here, we briefly present the performance of supervised approach using the train dataset since we cannot evaluate the accuracy of test dataset at current stage. We used 5-fold cross-validation using labeled data in train dataset.

Accuracy of First Phase. The first phase focuses on characterizing of home and workplace among frequently visited places. We found that mobility pattern (i.e., residence-time distribution at weekday and weekend) and environmental features are sufficient to infer home and workplace during the first phase. We set the threshold of visit frequency as 0.3 to extract frequented places (see Figure 2(a)). The first phase chose 139 places from labeled data, and it contains 71 homes and 59 workplaces, as shown in Table 3. It contains 6% of other places (i.e., 7 others' home, 1 outdoor sports, 1 shopping place), and recognized 2 places as home or workplace. The first phase correctly recognizes 94% of places since data at home and workplace include unique features.

Accuracy of Second Phase. The second phases infer 10 categories within minor places and places recognized as 'others' at first phase. In the second phase, we eliminated the residence-time distribution at weekend since it does not show unique characteristics across different type-of-places. Table 4 present the accuracy of second phase. Among 197 places, 97 place are correctly recognized and the overall accuracy combining the result of first phase is 68%. Considering the number of classes, the proposed method shows high accuracy, but the method is still limited for differentiating similar type-of-places (e.g., my home/others' home or my workplace/others' workplace) or minor places, as shown in Table 4. Additionally, we expect that the accuracy of test dataset would be decrease since most places of labeled data in train set are home and workplace (i.e., 130 of 336 places).

4. CONCLUSION

In this paper, we presented topic-modeling based place characterzing method to infer type-of-places. We analyzed various features in data collected by smartphone users, and explored behavior patterns and environmental features which are useful to differentiate type-of-places. The method is built around a topic model based approach to place characterizing. Our results showed that the proposed method able to automatically classify places into the 10 different categories with an accuracy of 68%.

Table 4: Confusion matrix at second phase.

Result Ground truth	My home	Others' home	My workplace	Transportation	Others' workplace	Outdoor sports	Indoor sports	Restaurant / bar	Shopping place	Vacation spot
My home	0.73	0.10	0.00	0.00	0.00	0.03	0.00	0.07	0.03	0.03
Others' home	0.37	0.26	0.03	0.00	0.00	0.14	0.06	0.14	0.00	0.00
My workplace	0.07	0.00	0.71	0.00	0.13	0.00	0.04	0.05	0.00	0.00
Transportation	0.00	0.00	0.00	0.40	0.07	0.13	0.00	0.13	0.27	0.00
Others' workplace	0.00	0.00	0.20	0.40	0.40	0.00	0.00	0.00	0.00	0.00
Outdoor sports	0.00	0.00	0.05	0.15	0.00	0.25	0.25	0.20	0.05	0.05
Indoor sports	0.00	0.00	0.00	0.00	0.10	0.20	0.50	0.10	0.10	0.00
Restaurant/bar	0.00	0.00	0.10	0.00	0.10	0.20	0.10	0.50	0.00	0.00
Shopping place	0.00	0.00	0.07	0.00	0.20	0.13	0.07	0.27	0.27	0.00
Vacation spot	0.33	0.00	0.00	0.33	0.00	0.33	0.00	0.00	0.00	0.00

5. ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea grant funded by the Korean government, Ministry of Education, Science, and Technology (No.2011-0006464, No. 2011-0015332).

6. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[2] Y. Chon, N. D. Lane, F. Li, H. Cha, and F. Zhao. Automatically characterizing places with opportunistic crowdsensing using smartphones. In *Proc. 14th Int. Conf. Ubiquitous Computing (UbiComp'12)*. ACM, 2012.

[3] Y. Chon, H. Shin, E. Talipov, and H. Cha. Evaluating mobility models for temporal prediction with high-granularity mobility data. In *Proc. 10th IEEE Int. Conf. Pervasive Computing and Communications (PerCom'12)*, pages 206 –212. IEEE, 2012.

[4] K. Farrahi and D. Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Trans. Intell. Syst. Technol.*, 2(1):1–27, 2011.

[5] S. Isaacman et al. Identifying important places in people's lives from cellular network data. In *Proc. 9th Int. Conf. Pervasive Computing (Pervasive'11)*, pages 133–151. Springer-Verlag, 2011.

[6] J. K. Laurila et al. The mobile data challenge: Big data for mobile computing research. In *Mobile Data Challenge by Nokia Workshop, in Conjunction with Pervasive'12*, 2012.

[7] D. Ramage et al. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP'09)*, pages 248–256. ACL, 2009.

[8] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[9] P. Viswanathan, T. Southey, J. Little, and A. Mackworth. Place classification using visual object categorization and global information. In *Proc. 8th Canadian Conf. Computer and Robot Vision (CRV'11)*, pages 1 –7, 2011.

[10] L. Vu, Q. Do, and K. Nahrstedt. Jyotish: A novel framework for constructing predictive model of people movement from joint wifi/ bluetooth trace. In *Proc. 9th IEEE Int. Conf. Pervasive Computing and Communications (PerCom'11)*, pages 54–62. IEEE, 2011.

[11] M. Ye et al. On the semantic annotation of places in location-based social networks. In *Proc. 17th Int. Conf. ACM SIGKDD Knowledge Discovery and Data Mining (KDD'11)*, pages 520–528. ACM, 2011.