A Data Cleaning Framework for Enabling User Preference Profiling through Wi-Fi Logs

Yao-Chung Fan, Yu-Chi Chen, Kuan-Chieh Tung, Kuo-Chen Wu, Arbee L.P. Chen

Abstract—Nowadays mobile devices have become a ubiquitous medium supporting various forms of functionality and are widely accepted for commons. In this study, we investigate using Wi-Fi logs from a mobile device to discover user preferences. The core ideas are two folds. First, every Wi-Fi access point is with a network name, normally a human-readable string, called SSID (Service Set Identifier). Since SSIDs are often with semantics, from which we can infer the place where the user stayed. Second, a Wi-Fi log is produced when the user is near a Wi-Fi access point. A high frequency of a consecutively observed SSID implies a long stay duration at a place. To the best of our knowledge, our work is the first attempting to understand users from the collected Wi-Fi logs from mobile devices. However, Wi-Fi logs are essentially of various information types and with noises. How to assess the information types, eliminate irrelevant information, and clean up the noises within partial-informative SSIDs are therefore keys for profiling user preferences over Wi-Fi logs. In this paper, we propose a data cleaning and information enrichment framework for enabling the user preference understanding through collected Wi-Fi logs, and introduce a data clean framework for cleaning, correcting, and refining Wi-Fi logs. In addition, a comprehensive experiment with data collected from users is made to verify the effectiveness of the proposed techniques for cleaning noisy Wi-Fi data for user preferences profiling. We believe that this work opens a new direction for understanding users from a different perspective, and we make available the code and the collected data set used in this study to encourage further research in this direction.

Index Terms—Mobile Data, Classification, Data Cleaning, Experiments, User Modeling

1 Introduction

Understanding users can be a key for many business applications, such as recommendations making, categorical advertisements delivering, and customized services providing. In the past, numerous research projects have been conducted to analyze user behaviors, including web browsing, customer transactions, questionnaire investigation, social platform and media, etc. Over the recent years mobile devices have become a ubiquitous medium supporting various forms of functionality and are widely accepted for commons. With this trend, mobile devices can be viewed in a novel perspective: a mobile device is not just a mini computer for the device holder, but a personal behavior observer providing data around the holder or generated by the holder. With this viewpoint, mining data generated from mobile phones has also become an active research direction [1][2].

In this paper we investigate using Wi-Fi logs from a mobile device to acquire user understanding. To the best of our knowledge, our work is the first attempting to understand users from the collected Wi-Fi logs from mobile devices. There are two primitive observations for understanding users through Wi-Fi logs. First, every Wi-Fi access point is with a Service Set IDentifier (SSID), which is a 32 byte string. The SSID of a Wi-Fi access point is normally a human-readable string and thus commonly referred to as the network name of a Wi-Fi network. The SSID is typically named by the user who sets up the Wi-Fi network. Therefore, SSIDs are often with semantics, for example, the Wi-Fi access point of National TsingHua University is named as NTHU-WiFi, from which we can infer the place where the user stayed. Second, a Wi-Fi SSID is produced when the user is near a Wi-Fi access point. A high frequency of a consecutively observed SSID implies a long stay duration at a place. By these two observations, we can use the SSID with semantics to infer the information such as user identity and user preference. For example, one may infer the occupation of a user from the places the user visited daily, e.g., a graduate student may go to his/her laboratory every weekday.

Aiming at the above opportunities, with the support of hTC academic research project [3], a group of users were recruited to provide the Wi-Fi data through the built-in apps in the distributed hTC smart phones. The app performs scanning available Wi-Fi signals at a time interval of 15 seconds and sends the obtained Wi-Fi logs (Time, SSID, BSSID (Basic Service Set Identification), and Level of the Wi-Fi access point signal) to a data store server, where raw data from all participants are stored. In Table 1, we show an example of the collected Wi-Fi observations for a user. There are 152555 records collected with 301224 different Wi-Fi BSSIDs being observed from 65 participants. The Wi-Fi data collection starts from Aug. 31, 2013 to May 1, 2014. The total storage space is 28.7 Gigabytes.

With the various potentials of mining Wi-Fi logs, in this study, we target at discovering user preferences from the collected Wi-Fi data. The core idea is that the types of places a user highly visited might reveal his/her preferences or interests. While the idea seems feasible, we encountered several challenges when we started the mining process. First, the SSID is typically a very short string, such as a shortened form of affiliation. In fact, the SSID itself may not be very informative. Sometimes we can guess the meaning

- *Yao-Chung Fan is with Department of Computer Science and Engineering, National Chung Hsing University, Taiwan. E-mail: yfan@nchu.edu.tw*
- *Yu-Chi Chen and Kuan-Chieh Tung is with Department of Computer Science, National Tsing Hua University, Taiwan.*
- *Kuo-Chen Wu is with HTC Corporation*
- *Arbee L.P. Chen is with Department of Computer Science, National Chengchi University, Taiwan E-mail: alpchen@cs.nccu.edu.tw (corresponding author)*

TABLE 1
A portion of the collected Wi-Fi data of a User:(Time, SSID, BSSID, and Level of the Wi-Fi access point signal)

Time	SSID	Level	BSSID
2013/12/14 02:26:54 PM	TWM WiFi Auto	-72	d8:c7:c8:79:cb:d2
2013/12/14 02:26:54 PM	Jennifer's AP	-72	5c:63:bf:c9:84:9a
2013/12/14 02:26:54 PM	andrew	-74	64:66:b3:4c:6b:80
2013/12/14 02:26:54 PM	SHOYO	-73	74:d0:2b:88:6d:1c
2013/12/14 02:26:54 PM	wenshan	-94	00:13:f7:1b:c8:63
2013/12/14 02:28:44 PM	Andyhome	-90	f8:d1:11:75:54:5a
2013/12/14 02:28:44 PM	MAOWLAN	-81	00:18:e7:cb:6a:6c
2013/12/14 02:28:44 PM	unilevel	-94	90:f6:52:3a:e8:a4
2013/12/14 02:28:44 PM	Simon	-94	20:cf:30:87:dd:3b
2013/12/14 02:28:44 PM	7-11 WiFi	-95	90:f6:52:45:0c:24
2013/12/14 02:28:44 PM	Pomelo	-96	0c:82:68:34:90:22
2013/12/14 02:28:44 PM	TINASONIC	-79	00:1f:c6:27:e9:ce
2013/12/14 02:28:44 PM	MuJaHomeAP	-87	00:24:a5:34:0f:86
2013/12/14 02:28:44 PM	Starbuck-Wif	-93	78:54:2e:2f:3e:d0

Fig. 1. Word Clouds formed by the Collected SSIDs

of an SSID, such as SSID "NTHU-wifi," but in most cases we cannot, such as SSID "pas36." Second, the information encoded behind a given SSID is of various information types, such as a store, an affiliation, or no semantics. Some are useful to the user preference profiling application, but most are not. How to filter out the irrelevant information is therefore a key to enable the user preference understanding through Wi-Fi logs. Third, from the collected data, we see that for just a user, there are tens of thousands of different SSIDs observed during a short six months of the data collection period. How to effectively refine the information from the huge amount of information is therefore critical. In this paper, we target at the observed challenges and propose a data cleaning and information refining framework for enabling the user preference profiling through Wi-Fi logs.

The contributions of this paper are four folds.

- To the best of our knowledge, our work is the first attempting to understand users from the collected Wi-Fi logs from mobile devices. Existing works for understanding mobile device users are mainly based on user-generated GPS trajectories, which suffers from power-consuming and in-door place positioning limitations.
- We propose a data cleaning and information enrichment framework for enabling the user preference understanding through Wi-Fi logs.
- Under the proposed framework, we introduce a series of techniques for cleaning and correcting SSIDs based on various features derived from collected Wi-Fi data.
- A comprehensive experiment with real data is made to verify the effectiveness of the proposed techniques for refining information from the collected Wi-Fi logs.

The rest of the paper is organized as follows. In Section 2, we introduce a naive scheme to quickly show the idea and the challenges before the user preference profiling through Wi-Fi logs. In Section 3, we formally define the problem and the goal of data cleaning process for profiling user preferences. In Section 4, a series of techniques are introduced for the purpose of cleaning and correcting the collected Wi-Fi logs. Section 5 presents the experimental evaluation results and demonstrates the effectiveness of the proposed techniques. In Section 6, we review the related work and discuss the position of this paper. Finally, Section 7 concludes the paper and provides a research road map for mining Wi-Fi data for understanding mobile device users.

2 Naive Preference Discovery Scheme

In this section, we first present a naive scheme for analyzing the collected SSID data for profiling user preferences, and highlight the challenges we encountered from the naive scheme. The naive scheme consists of three phases: the SSID semantic expansion phase (Section 2.1), the user profile construction phase (Section 2.2), and the preference discovery phase (Section 2.3). Finally, in Section 2.4, we discuss the problems of the naive preference discovery scheme observed from the initial experiment results, which motivate the design of the proposed framework.

2.1 SSID Semantic Enrichment

An SSID is typically a very short string, which can be less informative. A fundamental idea to our study is that we make use of the web search service API, such as the Google Web Search API, to enrich the semantics of the SSID. With the help of Google Web Search API, we can readily expand the meaning of a given SSID. For example, if the SSID "nthu" is input into the API, we can obtain web documents regarding National TsingHua University, and if the SSID "pa36" is emitted, we obtain web documents of a performing arts school at Taipei called Performing Arts School 36. Therefore, with the employment of the web search API, a short, abbreviated SSID string can be expanded into web documents, which should be more informative than the original SSIDs. In Table 2, we show an example of the expanded words from top-11 highly observed SSIDs in the SSID data set shown in Fig 1.

2.2 Profile Construction

Given a user's SSID data, the process of the user profile construction is as follows. First, we sort the SSIDs by the appearing frequencies, as a high frequency of the observed SSID implies a long stay duration at a place and should be more meaningful to the targeted user. One thing to mention is that it is impractical to expand all the SSIDs. The reasons are two folds. First, the observed SSIDs are numerous. As mentioned, tens of thousands of different SSIDs have been collected for a single user. If all SSIDs are expanded, a very huge profile will be produced, which turns out to be less informative. Second, not every SSID is with the same importance. Therefore, after the sorting process, top-k highly observed SSIDs are selected to iteratively go through semantic enhancement process. In this study, we make use of the Google Web Search API to expand the possible semantics behind the SSID. When an SSID is input into the Google Web Search API, the API will return a set of documents. We then process the returned web documents by tokenizing words, segmenting words when dealing with

TABLE 2
An Illustrating Example for User Profile Creation

SSID	Freq.	Expanded Words after Enrichment Process
WiFi	35%	Wireless, Network, Telecommunication, Type, Radio, Computer, Power Wire, Nodes, Device, Application, Combination
EDIMAX	10%	edimax, Product, Service, Marketing, Thank You, Hello, Quality, Management, Welcome, Device, Contact Information
iLove4G_G1	8%	Activity, Taipei Free, Internet, 512k, Slow Speed, Inquiry, Global, Mobile, Why, High Speed, Login
HiLife_3G	7%	Convenient, Store, Authorized Chain, Start, Office, Corporation, Service, Selection, HiLife, Insisting, Humanity, Friendly
Pas36-3F	7%	Performance, Art, Shows and Exhibition, Classroom, Evaluator, Entrance, Open Day, FAX, Address, Tel, Weekends, Holiday
SAPIDO_RB-1802	6%	Nothing
Taipei-Free	6%	Taipei, Baby, Hot Spot, Enterprizes, Authorization, Public Area, Outdoor, Free, Setup, Traffic, Internet, Location, Message
Library-Dlink	5%	Created, Style, Links, Library, Ethernet, Adapter, Driver, Ebook, Javascript, User, External
NCCUDIP	5%	Post List, Location, NCCU, DIP, Institute, Diplomacy Department, National ChengChi University, College, School, Student
Eslite-Lecture-Room	1%	Audio-visual Room, Eslite Life, Urban City, Culture, Window of Humanity, Classics, Italy, Design, Projector, Sofa, Stairs, Seats
Starbucks-WiFi	1%	Coffee, Corners, Cozy, Always, Enjoy, Atmosphere, Sun, Windows, Web Page, Sweet, Partners, wifi

Chinese texts, and removing stop words. After the web document processing, the processed tokens are accumulated into the user profile. In Table 3, we show the constructed user profiles for the SSID data in Figure 1. In this example, we select top-5 frequently observed SSIDs for expanding the semantics.

In summary, by expanding the SSIDs, we construct a profile for the user, which contains a set of weighted words considered to be relevant to a user. The profile will be represented as a vector which serves as a basis for further data processing, such as similarity measurement between two profiles. The formal definition of a user profile is as follows.

Definition 1: User Profile A user profile P contains a set of weighted words, where the words are weighted by the number of occurrences $\omega_{P,w}$ of word w in the profile and is represented by a vector $(\omega_{P,1}, \omega_{P,2}, ..., \omega_{P,t})$, where t is the total number of the words in the all profiles.

2.3 Preference Discovery

Once the profile is constructed for a user, one challenge is how the profile is linked to the user preference. In the following, we first describe two terms regarding the following discussion: the preference topics and the preference score for a user profile over some preference.

Definition 2: Preference Topic A preference topic T is a set of weighted words that is considered to be relevant to the preference topic. The words in a preference topic are weighted by tf-idf weighting scheme over all available preference topics. A preference topic is represented by $(\omega_{T,1}, \omega_{T,2}, ..., \omega_{T,t})$. Formally, given that there are M preference topics, the weight $\omega_{T,w}$ of a word w is computed by word frequency $f_{T,w}$ in the preference topic and the topic frequency tf_w of the word (the number of topics containing w) by the following equation:

$$\omega_{T,w} = f_{T,w} \times log\frac{M}{tf_w} \quad (1)$$

Definition 3: Preference Score $S_T(P)$ In this paper, the score of a learned user profile P over the given preference topic T is defined as follows:

$$S_T(P) = \frac{\sum_{\forall\omega} \omega_{T,w} \cdot \omega_{P,w}}{\sqrt{\sum_{\forall\omega} \omega_{T,w}^2}\sqrt{\sum_{\forall\omega} \omega_{P,w}^2}} \quad (2)$$

With the definitions, our idea of discovering user preferences is to compute preference scores over M preference topics as a judgement for preference understanding.

2.4 Discussion

By the initial study and experiment results, we observe that the resultant user profiles are not as informative as we expected to describe a user. As an example, consider the profile shown in Table 3. The profile contains many keywords about network devices, and network providers, which are unlikely to be relevant to a user. In fact, from the experiments, we find that most user profiles produced by the naive scheme are in this case, i.e., the user profiles are filled with irrelevant keywords, which lessens the descriptiveness of the profile. In fact, by further examining the SSID data, we see that many SSIDs are without semantics or named by the device default setting, such as "ZyXel." Including the expanded words from these SSIDs makes the resultant profile to be less informative to the user preference understanding. Therefore, before the SSID data can be mined for profiling users, one important step is to reduce the influence from the less informative SSIDs.

As an example, in Table 4, we show the user profile by manually removing the noise SSID, such as WiFi, Edimax, iLove4G, etc. From the profile in Table 4, one can observe the words in the profile, such as cafe, classics, elite life, arts, performances, and diplomacy department, are more likely to describe a user or his/her preference compared with the resultant profile in Table 3. However, manually examining over tens of thousands SSIDs are not possible. An automatic approach for cleaning, correcting, and assessing the information encoded behind SSIDs is needed before constructing user profiles.

Yet another observation is that highly observed SSIDs may not truly reflect the preference of a user. By our observation, the SSIDs with high frequencies are often with daily stayed places, and tend to reveal the identity of a user, such as nicknames and occupation. In fact, in comparison with high frequency ones, relatively low frequency SSIDs may reveal more about user preferences. For example, the SSIDs collected during a user's vacation will have relatively low frequencies when compared with daily observed places. However, the SSIDs might tell us more about the users. For example, if a user always lodged at luxury hotels during a vacation, then it is reasonable to infer that the user prefers luxury things. Therefore, selecting the SSIDs with high frequencies might not be an effective approach for user preference discovery. Therefore, a mechanism for effectively selecting SSIDs for profile construction is needed for future design.

3 Problem Formulation

We first define two terms for measuring the descriptiveness of a produced user profile.

TABLE 3
A Profile by Naive Scheme

Profiled by Wi-Fi, Edimax, iLove4G, HiLife_3G, Pas36-3F
Wireless, Network, Telecommunication, Type, Radio, Computer, Power Wire, Nodes, Device, Application, Combination Edimax, Product, Service, Marketing, Thank You, Hello, Quality, Management, Welcome, Device, Contact Information Activity, Taipei Free, Internet, 512k, Slow Speed, Inquiry, Global, Mobile, Why, High Speed, Login, Convenient Store, Authorized Chain, Start, Office, Corporation, Service, Selection, HiLife, Insisting, Humanity, Friendly Performance, Art, Shows and Exhibition, Classroom, Evaluator, Entrance, Open Day, FAX, Address, Tel, Weekends, Holiday

TABLE 4
A More Informative Profile

Profiled by Pas36-3F, NCCUDIP, Eslite-Lecture-Room, Wi-Fi Starbucks
Performance, Art, Shows and Exhibition, Classroom, Evaluator, Entrance, Open Day, FAX, Address, Tel, Weekends, Holiday Post List, Location, NCCU, DIP, Institute, Diplomacy Department, National ChengChi University, College, School, Student Audio-visual Room, Eslite Life, Urban City, Culture, Window of Humanity, Classics, Italy, Design, Projector, Sofa, Stairs, Seats Coffee, Corners, Cozy, Always, Enjoy, Atmosphere, Sun, Windows, Web Page, Sweet, Partners, wifi

In this study, we consider a highly descriptive user profile as one with a high degree of overlapping with the given set of topics. To measure the descriptiveness, we define *user profile utility* as follows.

Definition 3: User Profile Utility $\upsilon(P)$ The utility of a user profile is defined by the following measure over the given preference topics:

$$\upsilon(P) = \sum_{w \in P} (\omega_{P,w} \times \sum_{\forall T} \omega_{T,w}) \tag{3}$$

In this paper, we consider the size of user profiles be limited for two reasons. First, overlarge profile may not well represent a user profile. Second, overlarge profile may also cause the curse of dimensionality problem. In what follows, we introduce a parameter to constrain the size of the user profile.

Definition 4: User Profile Restriction β For each user, we have a collection of Wi-Fi SSIDs S. We use $\zeta(\cdot)$ to denote the result of the function that performs information enrichment for a given SSID s through a web search engine. For simplicity of discussion, we use $\zeta(s)$ to refer to the results of the information enrichment process. For a given number β, $\beta \in N^+$, the profile construction under the given parameter β is subject to a subset S_β of the original SSID set, where $|S_\beta| \leq \beta$, and the profile over the constraint is given by

$$P_\beta = \bigcup \zeta(s), s \in S_\beta \tag{4}$$

Problem Goal The goal of this paper is to find a user profile P_β that maximizes the user profile utility $\upsilon(P_\beta)$ over the given set of preference topics subject to a given number β for the profile construction. Formally, we have the following optimization formulation.

$$\begin{aligned} \textbf{maximize} \quad & \tfrac{1}{|P_\beta|} \textstyle\sum_{w \in P_\beta} (\omega_{P_\beta,w} \times \sum_{\forall T} \omega_{T,w}) \\ \textbf{subject to} \quad & P_\beta = \bigcup \zeta(s), s \in S_\beta \\ & |S_\beta| \leq \beta \\ & \beta \in N^+ \end{aligned} \tag{5}$$

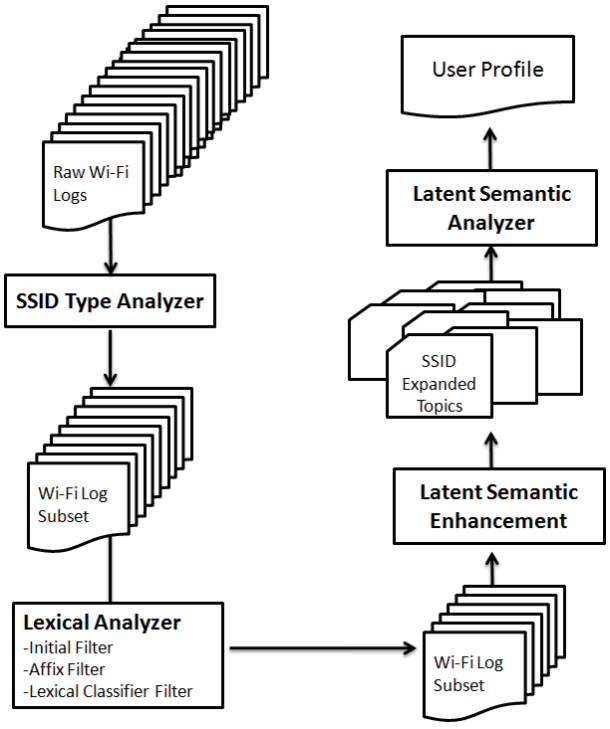

Fig. 2. The Flow Overview of the Proposed Framework

The proposed problem can be formulated as a weighted maximum coverage problem by considering the expanded documents of an SSID as a set of words, and the expanded result of all the SSIDs form a collection of such set. With the collection S and a given number of β, the objective is to find a subset $S' \subseteq S$ such that the weight of the included words is maximized. Such problem formulation is the well-known maximum coverage problem, which is shown to be NP-hard. Greedily choosing at each step is the best-possible polynomial time approximation algorithm.

4 The Proposed Cleaning Framework

Figure 2 shows an overview for the proposed cleaning framework. There are four key components in the proposed data cleaning framework: (1) SSID Type Analyzer, (2) Lexical Analyzer, (3) SSID Latent Semantic Enrichment, and (4) Semantic Analyzer. We discuss and elaborate these components in the following subsections.

4.1 Type Analyzer

As discussed, numerous SSIDs are observed and collected by a user. Including all SSIDs for profile generation does not make sense. A very fundamental step before mining the user preference is to select a quality subset of SSIDs as a basis for generating user profiles. Selecting by observing frequency is one possible solution. However, as shown in the naive scheme, selecting high frequently observed SSIDs is not an effective one; insteadly, high frequently observed SSIDs tend to reveal the identity of users, such as where he/she works or the nickname of the users.

In this study, we propose to incorporate an *SSID Type Analyzer* to select a subset of SSIDs according to the types of SSIDs. The main idea is two folds. First, during the data collection process, in addition to SSIDs, we also record the

time an SSID is observed. And, the time an SSID is observed is an effective indicator for making decision on inclusion of the examined SSID for profile generation. For example, one can select SSIDs observed in the weekend for user profile generation, as the SSIDs observed in the weekend are much likely to be the places for recreation and entertainment, and therefore much relevant to the user preferences. Similarly, one can select the SSIDs observed at night and noon for understanding the dinning preferences.

The idea of selecting SSIDs based on temporal feature can be further generalized and enhanced. Our idea is as follows. First, the location a Wi-Fi access point is installed is with a semantic type. For example, some Wi-Fi access points are installed at private location, such as home or personal offices, belonging to private location type, and some are installed at stores or restaurants, which are shopping type or catering type locations. And, the types of SSIDs (the location type that a Wi-Fi access point is installed) will be a good indicator for selecting a quality subset of SSIDs for user profile generation of various goal. For example, if the goal is to discover the identity of a user, then one can select the SSIDs that belongs to the private location type and working location type, and if the goal is to discover the preferences, then the SSIDs of store-type or catering-type locations should be included for profile generation.

However, the challenge for the above enhancement is that we have no information about the type of a given SSID. To this challenge, our idea is that human behaviors are not random but with regularity, e.g., people visit restaurants around noon, go for work in the daytime, and stay at home at night. Such regularity reveals the possibility of using the visiting patterns to infer the type of a location and the associated SSID. That is, we can cast the location type discovery problem as a classification problem and employ machine learning techniques to automatically infer the types of the SSIDs. In this study, we make use of the algorithm proposed [4], which capture place type characteristics by following four sets of temporal features (1) daily visiting pattern, (2) weekday visiting pattern, (3) stay duration visiting pattern, and (4) visiting frequency pattern to map the SSID into five types (working location, private location, catering location, shopping location, and recreation location).

In summary, given a set of user's Wi-Fi logs, the user profile is constructed as follows. First, the Wi-Fi logs are passed to the SSID type analyzer . The SSID type analyzer selects a subset of SSIDs based on their types. Note that the selection is based on the goal of the user profile. For example, if the goal is to discover the working affiliation, the SSID type analyzer will select the SSIDs that belong to the working type location. In the same manner, if the goal is to discover the preferences, the SSIDs belonging to the catering place and the recreation place will be selected to enter the next phase for the profile generation.

4.2 Lexical Analyzer

As discussed in the previous section, the other problem with using SSIDs for profile generation is that not every SSIDs are informative; some are without any semantics, and some with semantic but are less informative with respect to the user preference understanding. The SSID without semantics comes from that many Wi-Fi access points are named by meaningless characters, such as "ABCDEFG", "P888", and "Y036678", from which nothing can be derived. And the SSIDs with semantics but not informative come from that (1) the SSID named with the default SSID setting, which is a name given by equipment manufacturers, such as ZyXel, and Dlink, or a name set by network infrastructure providers, such CHT and iTaiwan, (2) some Wi-Fi access points are named by an affiliation or the owner name. Examples for affiliation type name are "nthu-cs", "nccu-dip", and etc. Examples for owner name type SSIDs are "yfan", "sam-lu", "Chou", and etc. While these SSIDs are with semantics, they are less informative in terms of user preferences understanding. In fact, the two types tend to reveal the identity of a user not the preference.

For the SSID types without useful semantics, such as device default name type and affiliation name type, the only thing we can do is to eliminate them from the given SSID set, as nothing can be derived from them. Therefore, a straightforward idea is to manually enumerate SSID names obviously without useful semantics and then filter them out during the profile construction process. To this end, we order all collected SSIDs from all participants by their appearing frequency and manually select highly observed SSIDs that are considered to be obviously without any useful semantics to the application. The SSID names, such as ZyXel, DLink, and Asus, are examples to be included in the list, as they are without any other useful information can be derived.

However, one point to mention is that it is impossible for the list-based approach to be effective. The major reason is that there are too many to list them all. As mentioned, there are 301224 different SSIDs observed in our project, which makes manually listing to be impractical.

For the purpose of judging the informativeness of an SSID from the lexical level, we observe the following clues. First, SSIDs are named by humans and often show language features on the given strings. We observe that SSID strings often contain delimiter characters, e.g., hyphen, whitespace, and underline. The delimiters can be used to chop an SSID string into tokens. The basic observation is that an SSID with many tokens is often informative to the places where the Wi-Fi access point is installed, e.g., "nthu-cs-7f-1" and "GRACE BOUTIQUE CAFE." In addition, there are some other features that can be employed. For example, if an SSID is with all upper-case letters, it's likely that the SSID is an abbreviation of something, such as a place or an affiliation, e.g. NTHU. Yet another observation is that if an SSID is with many digits, the SSID probably is without too much semantics. Therefore, to leverage these characteristics, the idea is to make use of supervised machine learning approaches on the basis of a training data set to have a binary classifier to judge if a given SSID is without semantics.

Therefore, for a given SSID, we compute the following features for the SSID: (1) the number of tokens, (2) the average token length, (3) the number of delimiters, (4) the number of digits, (5) the number of upper-case letters, and (6) the number of lower-case letters. As an example, for an SSID "nthu-MAKE_Lab sam38" we can extract the features from the SSID into the following form: [4, 4, 3, 2, 5, 9]. With the features, we can adopt supervised machine learning techniques to judge if the information encoded in an SSID is informative.

In summary, after the filtration of the SSID type analyzer, the reminder of the SSIDs are first sorted by the appearing

frequency. And then, the SSIDs are examined based on the frequencies from high to low until the restriction β is reached. For each SSID s under examination, we make the decision on whether to include it as the basis for profile generation based on the proposed lexical analyzer discussed above.

4.3 Latent Semantic Enrichment and Semantic Analyzer

In addition to the lexical level features of an SSID, another clue for the informativeness judgment is to leverage the semantic level features of an SSID. As aforementioned, by emitting an SSID into a web search engine, one can obtain a number of returned web documents which are considered to be relevant to the SSID. With the returned documents, the idea is to assess the informativeness by analyzing the contents of the documents.

As shown in the naive scheme, an SSID string is without much information, and the external knowledge, such as web search API, is employed to enrich the semantics of the SSID. Therefore, a naive idea is to include all the expanded information of the SSIDs selected by the two previous techniques as the user profile. However, one thing to point out is that not every SSID reached this stage is informative to preference understanding. In the two previous phase, the selection of SSIDs is mainly based on the type of an SSID and the lexical-level features of an SSID. The type of the SSID is inferred by the temporal patterns (the time the SSID is observed). The lexeme of an SSID with types relevant to preference is not necessary to be informative. For example, it is not uncommon to see that an SSID with catering type is with the device default SSID name. And, for the lexical analyzer, its goal is to positively eliminate the SSIDs obviously without information. In fact, we have no idea about the informativeness of most SSIDs at this stage.

A straightforward idea toward this informativeness assessment problem is to make use of supervised machine learning approaches on the basis of a training data set to have a binary classifier to judge if a given SSID and its expanded words is informative to the preference understanding and should be included into the profile being constructed. However, from the initial experiment result, we observe such idea is not effective in terms of producing quality profiles. The problem is that incorporating the information from web documents indeed enhanced the semantic but it also incurs another kinds of noise into the profile being constructed. Notice that for an SSID, we obtain a set of web documents from web search API. The web documents are treated as the expanded semantic for the SSID. However, problems with the idea is that (1) not every web document are informative to user preference, and (2) even a web document is informative, not every word in the document is relevant to the user preference, and directly including all the words containing in the document lessen the descriptiveness of the constructed profile. The idea of incorporating a classifier to asses the informativeness of a document may address the first problem, but not for the second one, as not all words in the document are relevant to user preference understanding. That is, a classifier at document level will not be effective in producing high quality profiles.

To address this issue, our idea is to apply Latent Dirichlet Allocation (LDA) [5] to discover word correlations at latent semantic level, and construct classifier at the level of topics rather than at document level.

LDA is a generative probabilistic model, and often used to discover latent topics in a given set of documents and the words associated with each topic. Given a set of documents $\{d_1, ..., d_m\}$ and a predefined number of latent topics Z, the LDA topic modeling process can be viewed as finding a mixture of topics for each document d_i, i.e., $P(z_j|d_i)$, with each topic z_j described by terms following an unobserved term-topic affiliation distribution, i.e., $P(t_k|z_j)$. Formally, the generative process can be formalized as

$$P(t_k|d_i) = \sum_{j=1}^{Z} P(t_k|z_j)P(z_j|d_i) \tag{6}$$

, where $P(t_k|z_j)$ is the probability of t_k belonging to topic z_j and $P(z_j|d_i)$ is the probability of picking a term from topic z_j for d_i. With the formulation, one can estimate the topic-term probability $P(t_k|z_j)$ and the document-topic probability $P(z_j|d_i)$ from the given set of documents by applying Gibb sampling [6] based on the Dirichlet priors.

We collect all the expanded results $\{\zeta(s)|s \in S_\beta\}$ from the set S_β of SSIDs passed from the lexical analyzer, perform a preprocessing by tokenizing words, segmenting words when dealing with Chinese texts, and removing stop words, and then apply the LDA technique over the set of preprocessed documents. LDA assigns to each term latent topics together with a probability describing the the strength the term related to a latent topic. For each latent topic we select top-100 terms to describe the topic. For now, each topic contains a set of words considered to be associated with the topic.

One thing to mention is that not every topic is related to the preference understanding; some are related to other issues, and some cannot be interpreted. Therefore, with the discovered topics, our idea is to make use of supervised machine learning approaches on the basis of a training data set to have a binary classifier to judge if a topic is related to preference understanding and should be included into the profile being constructed. The construction of the classifier is as follows. First, we expand all SSIDs from all the participants and again apply LDA to discover topics. After that each topic keeps the 100 words most strongly associated with that topic to represent the topic. Then, three experts wade through all the topics and manually label the topics into preference relevant and preference non-relevant. The labeled topics are then as a training data from which a binary classifier is learned.

As a summary of this stage, this component takes the SSIDs passed from the type analyzer and the lexical analyzer, and start from the dimension of the expanded web content to examine if an SSID is informative to user preference understanding. More specifically, the SSIDs at this stage are expanded into sets of documents, and then the results are went through the topic modeling process to discover latent topics between words. Finally, a binary classifier is employed to make decision on including or excluding the produced topics.

5 Performance Evaluation

In this section, we provide the experiment results. In Subsection 5.1, we describe the experiment settings and the data set

for evaluation. Subsection 5.2 presents the experiment result for a comparison overview and Subsection 5.3 discuss the results.

5.1 Experiment Settings

5.1.1 Raw Wi-Fi Data

With the support of hTC academic research project, in our study, 65 participants are recruited by giving the free new hTC smartphones in exchange for contributing all the smart phones usage data in two years through the built-in apps in the distributed hTC smart phones. The app performs scanning available Wi-Fi signals at a fixed time interval and sends the obtained Wi-Fi observations (Time, SSID, BSSID, and Level of Wi-Fi signal) to a data store server, where raw data from all participants are stored. The Wi-Fi data collection starts from 31 Aug. 2013 to 1 May 2014. We use the collected raw Wi-Fi data trace as data sets for experiments.

5.1.2 Preference Topics

In the experiments, we perform experiments with two preference topics sets: the one from Google AdWords, and the one generated from the Wi-Fi data set we collected.

- **Google AdWords Topics** we select three topics from Google AdWords: (1) restaurant and night life, (2) arts and entertainments, and (3) travel and sightseeing. In the topic of restaurant and night life, there are 4134 keywords (suggested by Google AdWords service), in the arts and entertainments, there are 5413 keywords, and in travel and sightseeing, there are 6749 keywords. We use these three topics and their associated keywords as topic preference profiles in the following experiment evaluation.
- **Wi-Fi Topics** In addition to the employment of Google AdWords preference topics, we also generate preference topics by applying the LDA technique over the expanded SSID documents from all collected SSIDs of all participants, and manually select topics related to (1) restaurant and night life, (2) arts and entertainments, and (3) travel and sightseeing, the same topics as we selected from Google AdWords, as the preference topics to evaluate the performance of the proposed scheme. In the topic of restaurant and night life, there are 200 keywords, in the arts and entertainments, there are 360 keywords, and in travel and sightseeing, there are 400 keywords.

5.1.3 Evaluation Metrics

Ground Truth As the goal is to discover the user preferences, in the experiment we ask the participant to rank the preference topics according to their favorite over the topics. We ask the participant to give a favorite order for the following three preference topics: (1) restaurant and night life, (2) arts and entertainments, and (3) travel and sightseeing. We use the this order as the ground truth to evaluate the performance of the proposed schemes.

Performance Metrics We perform the experiment evaluation by considering the following two performance metrics: the utility measure proposed in this paper and the Normalized Discounted Cumulative Gain (NDCG) measure, which is a measure for evaluating the effectiveness of ranking applications. In the experiments, we implement the proposed scheme to generate user profiles, and leverage the proposed preference scores to rank the user preferences among the given preference topics. With a computed preference order, we calculate the Discounted Cumulative Gain (DCG) for the preference order, and treat the preference order given by users as an ideal result to normalize the DCG of the preference order computed by the proposed preference discovery scheme. One thing to note is that if a computed preference score is smaller than 0.01, then the preference topic is excluded from the ranking due to statistical insignificant concern.

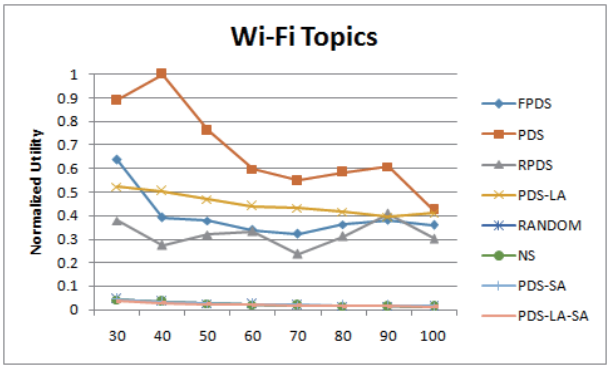

Fig. 3. Utility by Varying β with Wi-Fi topics data set

5.2 Performance Comparison

In this subsection, we provide experiment results over the given data sets to have a performance overview of the compared schemes. In the experiments, three elementary schemes are compared: NS, PDS, RANDOM. In addition, we also compare the other schemes with various analyzer combinations to observe the performance of individual analyzers.

- **NS** We implement the naive scheme called NS for generating profiles proposed in Section 2.
- **PDS** We implement the proposed preference discovery scheme PDS introduced in Section 4 for profile generation. PDS consists of the SSID type analyzer, the lexical analyzer, and the semantic analyzer.
- **RS** We also implement a random selection scheme called RS, where SSIDs are randomly selected for profile generation and directly expanded and includes as generated profiles without further any informativeness assessment.
- **FPDS** This scheme is a variant of PDS by replacing SSID type analyzer with a component that selects SSIDs based on the observed frequency.
- **RPDS** This scheme is a variant of PDS by replacing SSID type analyzer with a component that randomly selects SSIDs for profile generation.
- **PDS-LA** This scheme is a variant of PDS by removing the lexical analyzer from the implemented PDS.
- **PDS-SA** This scheme is a variant of PDS by removing the semantic analyzer from the implemented PDS.
- **PDS-LA-SA**This scheme is a variant of PDS by removing the lexical analyzer and the semantic analyzer from the implemented PDS.

5.3 Result Discussion

Figure 3 and 4 shows the experiment results over Adwords data set, and Figure 5 and 6 shows the experiment results

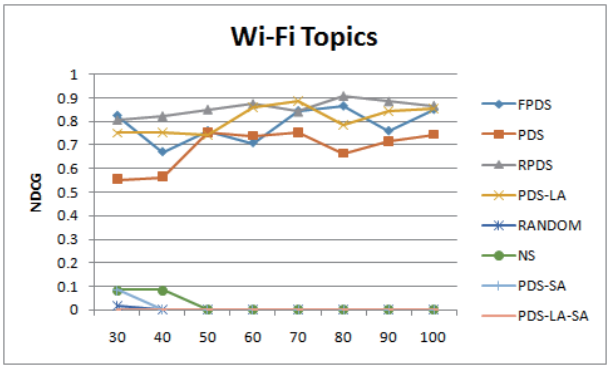

Fig. 4. NDCG by Varying β with Wi-Fi topics data set

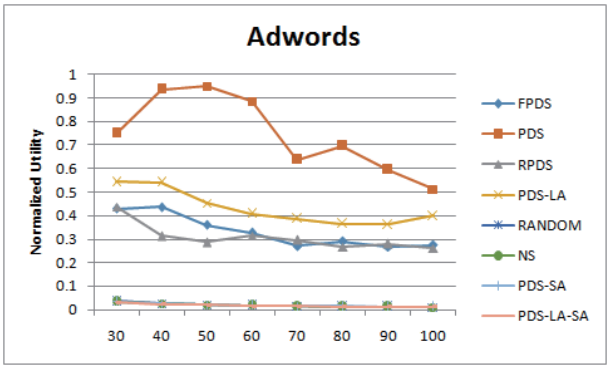

Fig. 5. Utility by Varying β with Adwords data set

over Wi-Fi topics data set, where x-axis is the value of parameter β, y-axis in Figure 3 and 5 is the averaged utility values for the compared schemes, and y-axis in Figure 4 and 6 is the averaged NDCG measure for the compared schemes.

We have four observations for the experiment results. First, in terms of utility metric, we observe that PDS consistently outperforms the other schemes along the varying of the β values. The performance trends are similar with the tested two data sets; the NS and RANDOM scheme without incorporating any analyzers proposed in the study shows the worst results, the other schemes incorporating one or two of the proposed analyzers show different performance results (we discuss the implication behind the difference), and PDS shows the best results.

Second, the semantic analyzer plays an important role for the quality of the generated profiles. One can observe that the schemes without incorporating the semantic analyzer, i.e., NS, RANDOM, PDS-SA, and PDS-LA-SA, are all with very low utility and NDCG values. This result can be expected, as we enrich the semantics of the SSID strings by search engines, and not every expanded SSID documents are relevant to user preference understanding. Without the helps from the semantic analyzer, directly including the expanded documents of the SSIDs passed from the lexical analyzer will result in less descriptive profiles, as the lexical analyzer aims to rule out the SSIDs obviously without semantics, such as an SSID with a sequence of numbers.

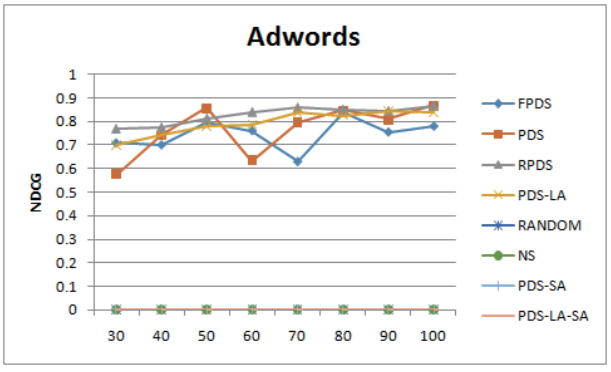

Fig. 6. NDCG by Varying β with Adwords data set

Third, when RPDS, FPDS, and PDS are compared, we observe that the incorporation of the SSID type analyzer is also critical to the quality of the generated profiles. RPDS, FPDS, and PDS are different only at the type analyzer stage, where RPDS select highly frequently observed SSIDs, RPDS randomly select SSIDs from raw data set, and PDS first select SSIDs according to the SSID types, and then the observing frequency. From Figure 3 and 5, we observe that there is a significant degrade between FPDS and PDS, which demonstrates the effectiveness of the idea of selecting SSIDs based on the SSID types. Also, there is another degrade between RPDS and FPDS, showing the value of selecting SSIDs by considering the factor of the observing frequency. One thing to note is that RPDS still provides mediocre results. This is because while the SSIDs are selected at random in RPDS, it still consists of the other two analyzers: lexical and semantic analyzer, which as discussed also contribute to the quality profile generation.

Forth, when PDS-LA, RPDS, PDS-SA are compared, we observe the fact that among the three proposed analyzers, the semantic analyzer is of the great importance, the SSID type analyzer is the second, and the lexical analyzer is the last. In Figure 3 and 5, we observe that when the semantic analyzer is removed from PDS, i.e., PDS-SA scheme, there is a huge performance degrade between PDS and PDS-SA, showing the importance of the semantic analyzer. Similarly, when the lexical analyzer is removed from PDS, i.e., the PDS-LA scheme, the scheme still provides mediocre results with the support the semantic analyzer.

6 Related Work

Understanding users can be a key for many business applications. In the past, numerous research works have been conducted to analyze users through various dimensions. Recently, there are two active directions for understanding users from mobile device data or social media data. This section reviews the existing works and discusses the position of our work.

Mobile Device Data In this direction, user modeling is performed by the data collected or generated by mobile devices of users. The viewpoint is to treat a mobile device as a personal behavior observer providing data generated by the holder. Understanding users is therefore possible by properly mining the collected mobile device data. There are two main research projects on this direction: MircoSoft GeoLife [1] and MIT Reality Commons [2].

To understand the users from mobile device, one method is by user trajectories, which record users' location history, because the trajectories imply to some extent users' interests and preferences. Along this direction, the Microsoft Geolife

project [7][1][2][8] developed a series of techniques for enabling users understanding by GPS trajectories. However, the techniques based on GPS trajectories suffer from the following concerns. First, the GPS sensor is with high power consumption. It is unlikely for users to always keep GPS sensors on for their device. Second, people spend the majority of their time indoor. However, the GPS system has the problem for positioning indoor locations, which limits the efficacy of the developed techniques.

In comparison with using GPS trajectories, our study investigates using Wi-Fi access point, which now well cover the urban city, as position indicators, as it is with the advantages of being able to position indoor users and having low power consumption.

The basic idea of using Wi-Fi based positions is that we can scan the available Wi-Fi access point signals near a user's current location via mobile devices. The observation is that Wi-Fi access points are unlikely to be moved, and for the same place, we should get a similar set of available Wi-Fi observations. By recording the Wi-Fi observation at a place, we will be able to know that the user is near the place. Using Wi-Fi observations as position indicators has the following advantages. First, the power consumption of scanning Wi-Fi fingerprints is less expensive with respect to using GPS positioning sensors. Second, the Wi-Fi based approach will not suffer from the indoor locating problem in comparison with the solution using GPS positioning.

Among the Wi-Fi based localization applications, the researches [9][10][11] on making use of Wi-Fi signal strength variations to position users in a space, such as a room, are the most active direction and have been studied extensively. The main idea is to first construct a database that records radio signals of a targeted indoor environment during training phase. Then, during positioning phase, the system is to find Wi-Fi fingerprints in the database most similar to the current fingerprint to estimate the current position. The goal of the techniques is to provide very fine-grained positions for tracked user (typically are within a few meters of accuracy); the goal and the proposed techniques along the Wi-Fi based localization are different to our goal and applications by understanding mobile device users through Wi-Fi SSID logs.

In Reality Commons Project, MIT Human Dynamics Lab addressed some challenges on making smart phones the essential tool for conducting social science research and also for supporting mobile commerce with a solid social science foundation. One great challenge is the lack of enough data in the public domain for capturing the disparate facets of human behaviors and interactions. MIT Human Dynamics Lab has built an environment to collect data, and the data collected open to the public for various usages. In Reality Commons, one project titled Reality Mining [2] has a goal to explore the capabilities of the smart phones that enable social scientists to investigate human interactions beyond the traditional survey or simulation based methodologies. 100 MIT students and faculty members were given a Nokia 6600 smart phone pre-installed with several pieces of data collection software. For a period of nine months, various types of data including call logs, user locations, Bluetooth devices in proximity, application usages, and phone status (such as charging and idle) were collected. Analyzing the Reality Mining data sets, Eagle et al. [12] [13] showed that the decision of individuals to identify a person as a friend was significantly correlated with whether he/she was with the person after work hours and during weekends, as captured by the smart phones. They also demonstrated the periodicity of an individual's personal behavior, identified by individual whereabouts, and the individual's interaction with others, both captured by smart phones [2]. While mining various mobile devices data are studied in [2], understanding users based on Wi-Fi logs is remained untouched.

The research in [14] investigates mining web logs from the end of a Wi-Fi access point instead of mining observed SSID logs from users' mobile phone (like we proposed in this study). And the research in [15] reports an initial study that examines a database of over 5 million wireless access points collected by Skyhook Wireless. By analyzing the default naming behavior, the location changes of access points over time, and the density of access points, the investigation suggest that the Wi-Fi access point data provide a fertile ground for understanding the What, Where and Why of Wi-Fi access points. However, the study still focused on understanding the places where the Wi-Fi access points installed rather than the idea proposed in this paper that understand the mobile device users through the observed SSID logs.

Social Media Data Yet another active direction for understanding users is by taking advantages of the explosion of social media. The studies on this direction were making attempts to analyze the users by their posts or actions, such as tweeting or Facebook messages over social media. For example, in [16], the authors study classifying users into democrats, republicans, and Starbucks aficionados by the tweets posted by a user. In [17], the authors propose a framework for estimating a Twitter user's city-level location based only on the tweet contents of the user. The research in [18] further proposes a large-scale topic model to represent tweets users and performs user following recommendation. In addition to tweet content analysis, the study in [19] propose to infer the genders of users based on the movies reviews from IMDB. In fact, numerous research have been conducted for profiling users through social media, and still remains as the most active research area for user modeling. While the goal is similar, the direction based on social media analysis is orthogonal to our direction. We believe that our work opens a new direction for understanding user from a different perspective.

7 Conclusion and Future Work

Understanding users is a key for many business applications. In this paper, we propose to pursue user preference understanding by their Wi-Fi logs collected from their mobile devices. As shown, Wi-Fi data are essentially of various information types and with noises. The challenges lie in how to refine relevant information from noisy Wi-Fi data. Aiming at the challenges, this paper proposes a data cleaning and information enrichment framework for enabling user preference understanding through Wi-Fi logs, and introduces a series of techniques for cleaning, correcting, and refining Wi-Fi logs. A comprehensive experiment with real data collected from users is made to verify the effectiveness of the proposed techniques for cleaning noisy Wi-Fi data for user preference profiling. To the best of our knowledge, this work is the first to study user behavior understanding by

mining Wi-Fi logs. This work is a beginning of a series of studies on mining Wi-Fi logs. In the following, we describe the other research issues under our current investigation.

During the data collection, in addition to the information of Wi-Fi access points, we also record the time the Wi-Fi information was observed. The logs are essentially sequential data ordered in timestamps. Therefore, sequential and cyclic patterns from the data can be discovered to understand the sequential and cyclic user behaviors. Having such information may have significant applications for recommendation services or mobile device resource managements. Furthermore, by properly reorganizing the Wi-Fi logs from different users, we will be able to discover co-appearances to a Wi-Fi access point. With this information, we can further find out the duration and times of the coincident appearance between two users. We can assume that two users with long coincident appearances have a special social relationship. Furthermore, if we can further investigate the type of the places and the time the two users met, we can further infer whether they are classmates, laboratory mates, or roommates. To our best knowledge, only few works [20] address this issue and the proposed approach is based on blue-tooth co-appearances, which may be unrealistic for practical use..

In addition, as discussed in the previous section, not every SSID is semantic-informative. Many Wi-Fi access points are named by default settings or without any semantics. However, human behaviors are not random, e.g., people visit restaurants around noon, go for work in the daytime, and stay at home at night. Namely, we can make use of the visiting patterns of the users to a place to infer the type of the place. With the collected user Wi-Fi logs and proper machine learning techniques, we can annotate the types of the places from the SSIDs without semantics. As aforementioned the Wi-Fi logs are sequential data, by gathering all users' Wi-Fi logs, we will be able to mine global sequential patterns, which help to understand global behaviors of users and to define the place relationships. Having such a place relationship, we will be able to detect popular travel sequences or the stream flow of people in buildings, which are valuable to interesting place recommendations and building managements, as the investigation [21] based on GPS trajectories.

Finally, we can view users with a mobile device as mobile sensors, and view a set of users as a mobile sensor network. The mobile sensor network moves with the users and provides various observations over the environment in a ubiquitous and real time manner. An interesting application from this viewpoint is that by collecting Wi-Fi logs for a sufficient time, we might be able to detect real-time events or anomalies for a place. The idea is that by analyzing the Wi-Fi logs, we can build a temporal periodic model for a place. A periodic model at a timeslot for a place records the number of users near the place for the timeslot. For example, about ten users at 3am stay at Place A. Having this information, we will be able to detect events or anomalies for Place A. For example, if hundreds of users were observed at the same timeslot, we know something unusual must have happened. The techniques for effectively defining the periodic models, compactly storing the models, and efficiently detecting the events are currently under our study.

REFERENCES

[1] Y. Zheng, X. Xie, and W.-Y. Ma, "Geolife: A collaborative social networking service among user, location and trajectory." *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, 2010.

[2] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and ubiquitous computing*, vol. 10, no. 4, pp. 255–268, 2006.

[3] "Htc corporation," http://research.htc.com/.

[4] C.-W. Chang, Y.-C. Fan, K.-C. Wu, and A. L. Chen, "On the semantic annotation of daily places: A machine-learning approach," in *Proceedings of the International Workshop on Location and the Web, LocWeb 2014, 2014*, 2014, p. 6. [Online]. Available: http://doi.acm.org/10.1145/1899662.1899668

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003. [Online]. Available: http://www.jmlr.org/papers/v3/blei03a.html

[6] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[7] Y. Zheng, Y. Chen, X. Xie, and W.-Y. Ma, "Geolife2. 0: a location-based social networking service," in *Mobile Data Management: Systems, Services and Middleware, 2009. MDM'09. Tenth International Conference on*. IEEE, 2009, pp. 357–358.

[8] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, "Cellular census: Explorations in urban data collection," *Pervasive Computing, IEEE*, vol. 6, no. 3, pp. 30–38, 2007.

[9] H. Shin, Y. Chon, and H. Cha, "Unsupervised construction of an indoor floor plan using a smartphone," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 42, no. 6, pp. 889–898, 2012.

[10] P. Prasithsangaree, P. Krishnamurthy, and P. K. Chrysanthis, "On indoor position location with wireless lans," in *PIMRC*, 2002, pp. 720–724.

[11] Q. Yang, S. J. Pan, and V. W. Zheng, "Estimating location using wi-fi," *IEEE Intelligent Systems*, vol. 23, no. 1, pp. 8–13, 2008. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/MIS.2008.4

[12] N. Eagle, A. Clauset, and J. A. Quinn, "Location segmentation, inference and prediction for anticipatory computing." in *AAAI Spring Symposium: Technosocial Predictive Analytics*, 2009, pp. 20–25.

[13] N. Eagle, Y. de Montjoye, and L. M. Bettencourt, "Community computing: Comparisons between rural and urban societies using mobile phone data," in *Computational Science and Engineering, 2009. CSE'09. International Conference on*, vol. 4. IEEE, 2009, pp. 144–150.

[14] D. Namiot, "On mining mobile users by monitoring logs," pp. 8–13, 2014.

[15] L. Liu, "What where wi: An analysis of millions of wi-fi access points," tech-report, Tech. Rep., 2007.

[16] M. Pennacchiotti and A.-M. Popescu, "Democrats, republicans and starbucks afficionados: user classification in twitter," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 430–438.

[17] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 759–768.

[18] D. Ramage, S. T. Dumais, and D. J. Liebling, "Characterizing microblogs with topic models." in *ICWSM*, 2010.

[19] J. Otterbacher, "Inferring gender of movie reviewers: exploiting writing style, content and metadata," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 369–378.

[20] B. Congleton and S. Nainwal, "Mining the mine exploratory social network analysis of the reality mining dataset," 2007.

[21] Y. Zheng, L. Liu, L. Wang, and X. Xie, "Learning transportation mode from raw gps data for geographic applications on the web," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 247–256.

ACKNOWLEDGMENTS

This research was supported by HTC Corporation and National Science Council, Taiwan.

Yao-Chung Fan received the PhD degree in computer science from National Tsing Hua University, Taiwan. His PhD dissertation received the 2011 Taiwan Institute of Electrical and Electronic Engineering Doctoral Dissertation Award and received the Taiwan National Science Council Award for visiting the Pennsylvania State University, U.S.A. He is currently an assistant professor in the Department of Computer Science at the National Chung Hsing University, Taiwan. His current research interests include mobile data mining, distributed data management, big data management, social computing, and sensor networks.

Yu-Chi Chen received the bachelor degree in computer science from National Taiwan Ocean University, Taiwan. She is currently a graduate student in the Department of Computer Science student at National Tsing Hua University, Taiwan. Her current research interests include mobile data mining, social media analysis, spatial-temporal pattern mining and uncertain pattern mining.

Kuan-Chieh Tung received the bachelor degree in computer science from National Chung Hsing University, Taiwan. He is currently a graduate student in the Department of Information Systems and Application at National Tsing Hua University, Taiwan. His current research interests include mobile data mining, social media analysis, information retrieval on microblog and big data management.

Kuo-Chen Wu received the master degee in degree in Bio-Medical Engineering from Chun-Yuan Christian University, Taiwan. He is leading Wearable Computing in HTC to focus on the technologies development of wearable devices, Bio-Signal process, Bioinformatics and system software.

Arbee L.P. Chen received the PhD degree in computer engineering from the University of Southern California, in 1984. He is currently the dean of the College of Science and University Chair Professor of computer science at National Chengchi University, Taiwan, R.O.C. He also holds a joint professorship at National Tsing Hua University. He was a member of Technical Staff at Bell Communications Research, New Jersey, from 1987 to 1990; an adjunct associate professor in the Department of Electrical Engineering and Computer Science, Polytechnic University, New York; and a research scientist at Unisys, California, from 1985 to 1986. His current research interests include data stream processing and analysis, data mining, and spatial databases. He organized the 1995 IEEE Data Engineering Conference and the 1999 International Conference on Database Systems for Advanced Applications, both held in Taiwan, and was a program committee cochair of the 2008 IEEE Data Engineering Conference held in Cancun, Mexico. He was invited to deliver a speech on music representation, indexing and retrieval at the US National Science Foundation (NSF)-sponsored Inaugural International Symposium on Music Information Retrieval at Plymouth, 2000, and was also invited to deliver a speech on searching music in a music collection in the IEEE Shannon Lecture Series at Stanford University, 2005. He has published more than 200 papers in renowned international journals and conference proceedings, and was a visiting scholar at Kyoto University, Japan, in 1999, Stanford University in 2003, 2004, and 2005, Kings College London, United Kingdom, in 2005, Boston University in 2009, and Harvard University in 2010 and 2011.