User Preference Profiling Through Wi-Fi Logs

S. Sobana
Student,M.tech(IT),
Dr. Sivanthi Aditanar College of Engineering,
Tiruchendur,Tamil Nadu

***Abstract*:- Nowadays, mobile devices have become a ubiquitous medium supporting various forms of functionality and are widely accepted for commons. In this paper, 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. Our work is the first attempting to understand users from the collected Wi-Fi logs from mobile devices. 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.**

Keywords: Data cleaning, Mobile applications, user profiling, Wi-Fi logs

1. INTRODUTION

Understanding users can be a key for many business applications, such as recommendations making, categorical advertisements delivering, and customized services providing. Over the recent years mobile devices have become a ubiquitous medium supporting various forms of functionality and are widely accepted for commons. In this paper, we investigate using Wi-Fi logs from a mobile device to acquire user understanding. Our work is the first attempting to understand users from the collected Wi-Fi logs from mobile devices.

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. SSIDs are often with semantics, For example, the Wi-Fi access point of National Chung Hsing University is named as NCHU Wi-Fi from which we can infer the place where the user stayed. 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 using this, we can 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. The core idea is that the types of places a user highly visited might reveal his/her preferences or interests.

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. Sometimes we can guess the meaning of an SSID, such as SSID "NTHU-wi-fi," 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, How to effectively refine the information from the huge amount of information is therefore critical. To solve these challenges, we propose a data cleaning and information refining framework for enabling the user preference profiling through Wi-Fi logs.

2. METHODOLOGY

We are using 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.

2.1 Type Analyzer

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, selecting high frequently observed SSIDs is not an effective one. High frequently observed SSIDs tend to reveal the identity of users, such as where he/she works or the nickname of the users.

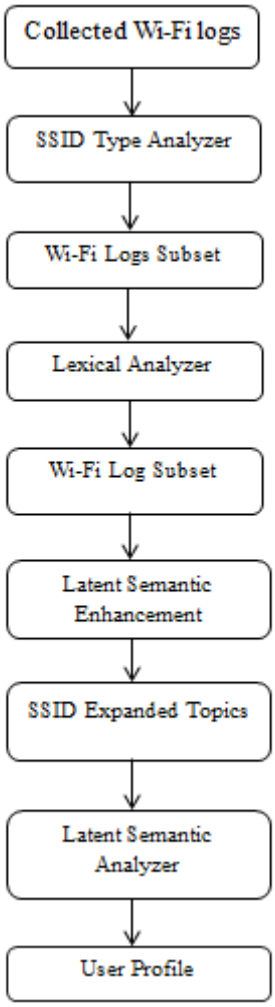

Fig. Flowchart of proposed framework

We propose to incorporate an SSID Type Analyzer to select a subset of SSIDs according to the types of SSIDs. 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 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. 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.

2.2 Lexical Analyzer

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 such as "nthu-cs", "nccudip" 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 onlything we can do is to eliminate them from the given SSID set as nothing can be derived from them.

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. After the filtration of the SSID type analyzer the reminder of the SSIDs are first sorted by the appearing frequency.

2.3 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. 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. 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. 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.

3. CONCLUSION AND FUTURE ENHANCEMENT

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 filters 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.

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 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.

4. 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, Apr. 2010.

[2] N. Eagle and A. Pentland, “Reality mining: Sensing complex social systems,” Personal and Ubiquitous Comput., vol. 10, no.4, pp. 255–268, 2006.

[3] 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 Proc. Int. Workshop Location and the Web, 2014, p. 6.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” J. Mach. Learning Res., vol. 3, pp. 993–1022, 2003.

[5] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” Proc. Nat. Academy of Sci. United States of Am., vol. 101, no. suppl. 1,pp. 5228–5235, 2004.

[6] Y. Zheng, Y. Chen, X. Xie, and W.-Y. Ma, “Geolife2. 0: A Locationbased social networking service,” in Proc. IEEE 10th Int. Conf. Mobile Data Manag.: Syst., Services and Middleware, 2009, pp. 357–358.

[7] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, “Cellular census: Explorations in urban data collection,” IEEE Pervasive Computing, vol. 6, no. 3, pp. 30–38, Jul.-Sept. 2007.

[8] N. Eagle, A. Clauset, and J. A. Quinn, “Location segmentation, inference and prediction for anticipatory computing,” in Proc. AAAI Spring Symp.: Technosocial Predictive Analytics, 2009, pp. 20–25.

[9] N. Eagle, Y. de Montjoye, and L. M. Bettencourt, “Community computing: Comparisons between rural and urban societies using mobile phone data,” in Proc. IEEE Int. Conf. Comput. Sci. and Eng., vol. 4, 2009, pp. 144–150.