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

(Extended Abstract)

Yao-Chung Fan[1], Yu-Chi Chen[2], Kuan-Chieh Tung[2], Kuo-Chen Wu[3], Arbee L.P. Chen[4]
[1]Department of Computer Science and Engineering, National Chung Hsing University, Taiwan
[2]Department of Computer Science, National Tsing Hua University, Taiwan
[4]Department of Computer Science and Information Engineering, Asia University, Taiwan
[3]HTC Corporation, Taiwan

I. Introduction

This paper investigates using Wi-Fi logs from a mobile device to acquire user understanding. There are two primitive observations for understanding users through Wi-Fi logs collected from mobile devices. 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 Chung Hsing University is named as NCHU-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 [1], 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 strength of the Wi-Fi access point signal). The data collection starts from Aug. 31, 2013 to May 1, 2014. There are 152555 records collected with 301224 different Wi-Fi BSSIDs being observed from 65 participants.

While the idea of acquiring user understanding via Wi-Fi logs seems feasible, we encountered the following challenges. First, the SSID is a short string, such as a shortened form of affiliation, which can be less informative. Second, the information encoded behind a given SSID is of various information types. Only some are useful to the user preference profiling. How to filter out the irrelevant information is therefore a key to enable the user preference understanding through Wi-Fi logs. Third, there are tens of thousands of different SSIDs observed during a short period. How to effectively select a quality subset of SSIDs for user understanding is therefore another critical issue. In this paper, we target at the observed challenges and propose a framework for enabling the user preference profiling through Wi-Fi logs.

II. Naive Preference Discovery Scheme

We first present a naive scheme for profiling user preferences through collected Wi-Fi logs, and discuss the problems with the naive scheme.

A. Profile Generation

Given a set of SSIDs observed by a user, the process of the user profile generation is as follows. First, we sort the SSIDs by the frequencies, as a highly frequently observed SSID implies a long stay duration at a place and should be more meaningful to the targeted user. Second, for a given SSID, we employ web search service API to enrich the semantics of the SSID. In such a manner, a short, abbreviated SSID string can be expanded into documents, which should be more informative than the raw SSIDs.

One thing to note is that it is impractical to expand all the SSIDs. The reasons are two folds. First, not all SSIDs are informative, as reported in [2]. Second, not every SSID is with the same importance. Therefore, after the sorting process, top-*k* high frequently observed SSIDs are selected to be expanded into a set of documents. When an SSID is input into a Web Search API, the API will return a set of documents. And, all words from the documents are treated as a bag of words and served as a profile for the user. The definition of a user profile is given 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.

B. 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, and then define 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

978-1-5090-2020-1/16/$31.00 © 2016 IEEE

preference topic and the topic frequency tf_w of the word (the number of topics containing w) by the following equation:

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

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:

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

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

C. Problem Formulation

We measure the descriptiveness of a user profile by *user profile utility*, which is defined as follows.

Definition 4: User Profile Utility $\upsilon(P)$ The utility of a user profile is defined by the following measure over the given preference topics:

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

Problem Goal The goal is to find a user profile that maximizes the user profile utility over the given set of preference topics subject to a given number β for the profile generation, where β is the number of SSIDs to be included for profile generation.

III. The Proposed Cleaning Framework

By the initial experiment results, we observe that the resultant user profiles of the naive user preference scheme are not as informative as we expected to describe a user. We find two problems for the naive scheme. First, highly observed SSIDs may not be relevant to the preference of a user. In fact, the SSIDs with high frequencies are often with daily stayed places, which tends to reveal the identity of a user. Second, not all SSIDs are with useful information with respect to user preference understanding. One obvious example is that many Wi-Fi aps are with SSID of device default setting, e.g. ”ZyXel”, which is nothing to do with user preference issues. Therefore, a mechanism for cleaning, refining, and assessing the information encoded behind SSIDs is needed.

There are three key components in the proposed data cleaning framework: (1) SSID Type Analyzer, (2) Lexical Analyzer, and (3) Latent Semantic Analyzer. The SSID type analyzer and lexical analyzer are focus on the issue of selecting a quality subset of SSIDs. The SSID type analyzer assesses the informativeness of SSIDs by considering the type information derived from the time an SSID is observed. The intuition is that the SSIDs observed in the weekend are much likely to be the places for recreation and entertainment, and therefore will be more relevant to the user preferences. And, the lexical analyzer is proposed based on the idea of judging the informativeness of an SSID from the lexical level. The intuition is that SSIDs are named by humans and often show language features on the given strings, which gives us clues to infer the informativeness of an SSID. On the other hand, the latent semantic analyzer advances the informativeness assessment of the expanded SSID content from the document level to topic level. The observation is that not every expanded document is informative to user preference, and even a document is informative, not every word in the document is relevant to the user preference, and directly including all words lessens the descriptiveness of the resultant profile. Therefore, we propose to first apply the topic model technique to form topics and perform informativeness assessment at the topic level rather than the document level.

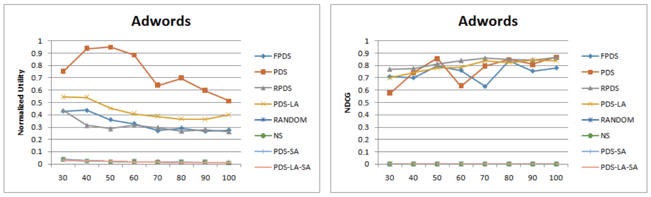

Fig. 1. Performance Comparison over Google Adword Topics

IV. Performance Evaluation

We perform the experiment evaluation by considering the following two performance metrics: the utility measure and the Normalized Discounted Cumulative Gain (NDCG) measure. The results are summarized in Figure 1. We observe that the performance of the proposed PDS scheme significantly outperforms the other schemes. Details about the experiment settings and other comparisons are reported in [2].

V. Conclusion

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. To the best of our knowledge, this work is the first attempt to study user behavior understanding by mining Wi-Fi logs.

Acknowledgment

This research was supported by HTC Corporation and Mini. of Science and Technology, Taiwan, under grant number 103-2221-E-005-051-MY2 and 103-2221-E-468-033-MY3.

References

[1] “Htc corporation,” http://research.htc.com/.

[2] Y. Fan, Y. Chen, K. Tung, K. Wu, and A. Chen, “A framework for enabling user preference profiling through wi-fi logs,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.