Detecting Wi-Fi Base Station Behavior Inappropriate for Positioning Method in Participatory Sensing Logs

Abstract
Recently mobile base stations are getting increased, which is considered harmful for the Wi-Fi positioning methods. In this paper, three approaches for detecting Wi-Fi base station behaviors inappropriate for Wi-Fi signature sampling are introduced and their performance evaluations are presented. First approach is for outdoor environment using GPS or Wi-Fi, second for indoor environment using Wi-Fi and accelerometers and last for the first contact stations using the Bayesian estimation method. Bayesian estimation is fine for stationary stations but much severe for mobile stations.

Author Keywords
Ubiquitous Computing, Sensing Network, Human Activity Recognition, Distributed Collaborative System, System Software, Embedded Software

ACM Classification Keywords
H.2.8 [DATABASE MANAGEMENT]: Database Applications- Data mining; I.5.1 [PATTERN RECOGNITION]: Models-Statistical; K.6.4 [MANAGEMENT OF COMPUTING AND INFORMATION SYSTEMS]: System Management- Quality assurance

Nobuhiko Nishio
Ritsumeikan University
1-1 Noji-Higashi
Kusatsu, Shiga 5258577 JAPAN
nishio@cs.ritsumei.ac.jp

Yuuki Fukuzaki
Ritsumeikan University
1-1 Noji-Higashi
Kusatsu, Shiga 5258577 JAPAN
yuuki@ubi.cs.ritsumei.ac.jp

Takuya Azumi
Ritsumeikan University
1-1 Noji-Higashi
Kusatsu, Shiga 5258577 JAPAN
takuya@cs.ritsumei.ac.jp
General Terms
Experimentation, Management, Measurement, Performance, Reliability

Introduction
Location sensing technology is very popular and useful for context aware service construction. Above all, GPS has gotten high availability since smart phones have prevailed. In the GPS-disabled areas such like indoor spaces, several other positioning technologies could be applied, where the Wi-Fi base station positioning is another popular method. Most of Wi-Fi positioning methods [1],[2],[3],[4] heavily rely on stationary installment of base stations, which means each base station is stationary and always turned on. However, the actual situation is quite different from this assumption. There exist many mobile base stations, since the tethering functionality of smart phones has been introduced. Some base stations are turned on at specific time periods, and to make matters worse, changes of people’s sitting locations or gathering situation are influential enough in observing fluctuating and unstable Wi-Fi radio signal status. Detecting such ill-behaving or should-not-be-used base stations is getting much more important than ever to keep positioning quality of Wi-Fi methods healthy.

The authors have been conducting participatory sensing project using Android smart phones for several years. This project has archived each personal sensor logs for these years through our participatory sensing platform. Archived sensing logs include GPS data, Wi-Fi base station observations, and other sensors log like accelerometer, compass, barometer and so on. These fragmented data logs are gathered and investigated in many ways. In this paper, we utilize logged data from participatory sensing of GPS, Wi-Fi base station signals and accelerometers and present three approaches to detect ill-behaving base stations described above, each with performance evaluation results. The first approach is using GPS log together with Wi-Fi observation log, which could be effective mainly in outdoor environment, and most Wi-Fi positioning services like Android location library might be applying for its convenience. The second approach is for indoor environment without GPS logs. The third approach adopts Bayesian estimation method.

In the rest of this paper, each of our three approaches is explained and performance results are given respectively. In the last section conclusion and future research direction are given.

Absolute Location Paired Approach for Outdoor Environment
Once Wi-Fi radio observation log is accompanied with absolute location information like GPS log at the same time, ill-behaving base stations like mobile Wi-Fi routers are relatively easily detected. From all the fragmented archives, only absolute location paired Wi-Fi logs are extracted and the geographical range of spanning for each unique base station is examined for verification of its stationary property. Standard deviation of observed absolute location logs is adopted for our approach and the formula to compute $S$ is in the following:

$$S = \sqrt{\frac{\sum_{i=1}^{n} ((\text{lat}_i - \text{lat}_{\text{ave}})^2 + (\text{lng}_i - \text{lng}_{\text{ave}})^2)}{2n}}$$
We use the Wi-Fi signal logs captured by our Android cell-phone-based sensing system for 121 days. While Wi-Fi signal logged every three seconds period and GPS logged every one second period, the most accurate GPS latitude and longitude data within ten seconds is paired with each unique Wi-Fi base station signal observation. There are 54 known mobile stations and 201 known stationary stations out of 14,963 unique stations. 2,446 stations cannot be paired with absolute location data because no GPS data is received. Figure 1 shows the known 54+201 stations downward sorted by standard deviation resulting maximum value of stationary station is 0.0094 and minimum value of mobile station is 0.0253. 0.02 would be fine for threshold value, although 16.3% of all stations could not be analyzed by this approach.

![SD in Location Samples](image)

**Figure 1.** Stationary/Mobile Wi-Fi Base Stations vs SD.

We also conducted the performance evaluation not using GPS but time sequence of Wi-Fi signal observation. If alternation of observed Wi-Fi signal statistics is recognized, we can know the observer’s movement. Figure 2 shows a sample observation of alternation of Wi-Fi signal observed in the train while GPS receiver badly worked. Here, as a threshold value for movement detection, we used eight consecutive scans that include unique Wi-Fi signal alternation. Out of 150 days log, we extracted Wi-Fi signal alternation (user’s movement) period paired without any GPS data, and conducted estimation for mobile base station. 214,053 unique base stations are observed and applying the movement recognition described above, the 51 stations are doubtful enough to be mobile base stations. We could also recognize that a few doubtful stations appeared from one train station and got off at the other station, since eight consecutive period unit is relatively short and efficient to know train stopping at the railroad stations.

![Wi-Fi Base stations which are observed four or fewer times](image)

**Figure 2.** Wi-Fi signal Alternation Observed.
Site Grouping Approach for Indoor Environment

In indoor environment, absolute location information is not automatically paired with Wi-Fi radio observation without any positioning method. We assume that our indoor daily activity could be considered as transition between frequently visited places. For example, at an office we have a place to work, desk, meeting room or café to take lunch. They might be changing from time to time, however they are not uncountable. We concentrate on such frequently visited and staying-long-time places while neglecting transitional states between them. These frequently visited places can be represented as Wi-Fi radio signal statistics when the accelerometer senses the user is not moving. Many of fragmented Wi-Fi signal log periods are recorded and they are analyzed in the light of similarity of radio signal statistics in order to make clusters. This similarity is calculated via the following formula. Here, \( F_w \) means a feature vector of base stations already sampled while \( F_c \) means the feature vector currently observed. \(| \text{vector} |\) means the number of base stations. Each of \( rssi \)'s in the equation is signal strength of each base station in the respective vector.

\[
dist(F_w, F_c) = \sqrt{\sum_{i} (rssi_{F_w} - rssi_{F_c})^2} \times \frac{|F_w| + |F_c| - |F_w \cap F_c|}{|F_w \cap F_c|}
\]

We introduce the "site" concept that is a cluster of log periods each of which have same base stations in common at certain ratio\(^3\). We assume that each site would represent relatively small range of space like narrower than the building floor and that each site includes a few places (that is represented as Wi-Fi signal statistics) where the user stay for a long period of time. We can concatenate every fragment of Wi-Fi logs that are observed at the same place as above so as to find ill-behaving base stations like mobile routers or temporarily turned-on/off usages through the accumulated long-term and same place observation log. Once the ill-behaving base stations are found or suspected, their information can be utilized at the same site in common.

\(^3\) We know that this includes a chicken-egg problem, but we optimistically consider it not so harmful. Mutual feedback might improve performance of both positioning and detecting.

\(^2\) [3] introduces other hierarchical grouping method but not for our objective.

---

**Figure 3. Clustering Places and Site Concept.**

Figure 3 introduces our site concept, and inside the same site ill-behaving station information from the long-term staying place Wi-Fi statistics can be shared by all other member places, while a few orphaned places remained to be poorly informed.
Figure 4. Site-Creation Experiment.

Experiment for site creation is conducted at our laboratory building. In Figure 4, four places around the laboratory are detected as staying places by accelerometer movement sensing, a created site is a black oval that holds three of places out of four.

Using the longest-term staying place statistics, concatenated observed result is shown in Figure 5. Each horizontal colored line represents one unique base station and they are upwardly sorted by their observation frequency. Simply thinking for avoidance of ill-behaving stations, the weight of each base station should be calculated using observation frequency like as follows:

\[
\text{weight} = \frac{\text{frequency}}{\text{maxFrequency}}
\]

However, there were some state-alternating stations pointed by solid arrow marks in Figure 5. These stations were well observed in some time, but suddenly changed in worse state in other time. Therefore, we revised the weight to show station’s illness as follows:

\[
\text{revisedWeight} = \frac{\text{frequency}}{\text{maxFrequency}} \times \frac{1}{1 + \text{LowFrequencyDuration}}
\]

The formula above adopts \text{LowFrequencyDuration} that is the longest duration when observation frequency was lower than a certain low limit threshold value.

Figure 5. Concatenated Wi-Fi Signal Observation Fragments.

Bayesian Estimation Approach for First Contact Base Stations

When encounter with first contact base station, the database does not have enough information to judge whether it is mobile base station or not. Therefore, we implemented a Bayesian Filter utilizing ESSID and BSSID to estimate mobile base station from our
Black/White list\(^3\) of archived base stations. The B/W list is created from the results of the absolute location paired approach to training a tendency whether the base station is mobile or not.

As for BSSID, 4 octets are used for training and estimation where vendor ID and device type are encoded. Parsing ESSID, only over three characters words delimited by special symbols or space characters are used. Such proper nouns like vendor or carrier names are manually picked up and used for training. Features extracted from BSSID and ESSID are in feature set \(F\), and \text{BLACK} or \text{WHITE} are in category set \(C\) in the following formula:

\[
p(C \mid F) = p(C) \prod_i p(F_i \mid C)
\]

\[
p(F_i \mid C) = \frac{T(C, F_i) + 1}{\sum_{F' \in V} T(C, F') + V}
\]

Here above, \(T(C, F)\) means number of feature \(F\) appearances when the category is \(C\), and \(V\) means total number of features in \(F\) when training phase.

\(^3\) Black means mobile and white means stationary here.
Figure 7. Bayesian Estimation Accuracy for Stationary and Mobile.

**Conclusion**

We conducted three approaches so as to detect Wi-Fi base stations behaving inappropriately for positioning methods. GPS data is used for absolute location information paired with Wi-Fi signal and applied in outdoor environment. In indoor environment, site concept is introduced to share ill-behaving base station statistics. After these analysis, training is done for the Bayesian filter to detect mobile base stations via BSSID/ESSID. Figure 8 shows these three approaches performance result for 150 days.

Our Bayesian estimation filter works fine for stationary stations but is rather much severe for mobile stations. In the future research direction, we should verify if our filtering technology would improve positioning method performance effectively.

**References**

