Bachelor's thesis

Machine Learning-based User Movement Prediction in Layer 2 Networks

Vorhersage von Benutzerbewegungen in Layer 2 Netzwerken basierend auf Maschinellem Lernen

by

Lina Wilske

Supervisors

Prof. Dr. Holger Karl

Leonard Paeleke

Internet Technology and Softwarization Group

Hasso Plattner Institute at University of Potsdam

July 29, 2023

Abstract

...

Contents

1 Introduction

In big Wireless Fidelity scenarios with multiple Access Points in office buildings, shopping malls, and airports people are moving around indoors with their mobile phones and need to be connected to an Access Point. This is called roaming and is a very important part of Wi-Fi, since 802.11k[2]. Roaming was improved with Access Point-initiated roaming, which is a feature of 802.11r[1]. But this roaming process does not include human movement, because if the Access Point1 knows, that the station is passing it and moves towards Access Point2, the Access Point2 should initiate the roaming process. This is not possible with the current roaming process, because the Access Point1 does not know, that the station is moving towards Access Point2.

Nowadays, using Machine learning to predict human movement in other scenarios like digital health or telecommunications is very common. A prediction of the next Access Point and initiate the roaming process before the station is moving towards the next Access Point? In this thesis, the prediction will be based on the Received Signal Strength Indications of the Basic Service Set Identifiers and waypoints of the station.

We need to decide, if we want to generate the data on our own or use existing data. We want to use machine learning to predict the next Access Point with data generated by users devices. Generating data on our own is not an option, because of the time and effort it would take to collect data. So we decided to use existing data. We will use a data set from kaggle[5] of a competition of Microsoft Research in 2021[3]. In this chapter, we will analyze the data. Furthermore, we will identify which parts of the data is needed for the machine learning model.

There are several machine learning models which could be used. We will take a look into some pre-selected models and discuss and propose a model, which could be the best for our data. This thesis will not concept a new machine learning model and also will not use combined models.

The proposed decisions will be implemented. As we need to preprocess the data for our model. We will prefilter specific parts of the data, because as we found out in, not all data is needed for the machine learning model. In this chapter, we will

preprocess the data and implement the machine learning model as well as discuss the chosen encryption. A big part of machine learning models is tweaking the hyperparameters.

At the end we will evaluate the model and discuss the results. We will also discuss, if the model could be used in a real-world scenario.

2 Data analysis

Data is the essential part of a machine learning model. Therefore, a wisely chosen dataset is needed. The dataset used in this thesis is the Indoor Location & Navigation from kaggle[5] which was part of a competition of Microsoft Research in 2021[3]. The data was recorded in shopping malls by a company called XYZ[10] and was provided by Microsoft Research for this competition. The goal for the competition was, given a site-path file, predict the floor and waypoint locations at a timestamp given in the submission files.

2.1 Components of the dataset

As noted in the kaggle notebook "Indoor Navigation: Complete Data Understanding"[4] the data consists of 3 parts:

- a train folder with train path files, organized by site and floor
- a test folder with test path files, organized by site and floor but without waypoint data
- a metadata folder with floor metadata, organized by site and floor, which includes floor images, further information and a geojson map

For this thesis, the submission files as well as the test folder will not be used, because our goal is another type of prediction. In the following, the train folder will be analyzed. In general every file was hashed, so that no specific information about the site is visible. The train folder consists of 204 subfolders, which represent each site where the data was recorded. In each site folder are a minimum of one and a maximum twelve subfolders, which represent the floors of the site, the median is 5 floors. Overall there are 26,925 files each representing a movement on a specific floor and site. Per floor, there are between one and 284 files with a median of 14 files. These files contain the information about the movement of a person on this specific site and floor. With this amount of data, it could be possible to train a machine learning model.

2.2 File structure

Each file contains the following information:

```
\#  startTime:1571462193934
\#  SiteID:5d27099303f801723c32364d SiteName: 银泰百货(庆春店) FloorId:5d27099303f801723c323650 FloorName:4F
1571462193944   TYPE_WAYPOINT   57.885998   69.501526
1571462194071   TYPE_ACCELEROMETER  -0.95254517 0.7944031   8.928757    2
1571462194071   TYPE_MAGNETIC_FIELD -25.65918   -4.4784546  -28.201294  3
1571462194071   TYPE_GYROSCOPE  -0.22373962 -0.07733154 -0.16847229 3
1571462194071   TYPE_ROTATION_VECTOR    0.04186145  -0.02101801 -0.72491926 3
1571462194071   TYPE_MAGNETIC_FIELD_UNCALIBRATED    -4.8568726  10.406494   -387.44965  20.802307
      14.884949   -359.24835  3
1571462194071   TYPE_GYROSCOPE_UNCALIBRATED -0.22218323 -0.068359375    -0.1628418  0.0026245117
      9.765625E-4 -7.6293945E-4   3
1571462194071   TYPE_ACCELEROMETER_UNCALIBRATED -0.95254517 0.7944031   8.928757    0.0 0.0 0.0 3
...
1571462194883   TYPE_WIFI   b06c4e327882fab58dfa93ea85ca373a54e887b5    9
      f967858afcbb907af6e5adef766c7e7b936ef07    -63 2462    1571462190744
1571462194883   TYPE_WIFI   8204870beb9d02995dab3f08aad97af5eab723cc    0413
      b35df78fc865af15b4721d5aeb33ff57da45    -64 2447    1571462188686
...
1571462194020   TYPE_BEACON 07efd69e3167537492f0ead89fb2779633b04949
      b6589fc6ab0dc82cf12099d1c2d40ab994e8410c    76e907e391ad1856762f70538b0fd13111ba68cd    -57 -71
      5.002991815535578   1b7e1594febd760b00f1a7984e470867616cee4e    1571462194020
...
\#  endTime:1571462195976
```

BSSIDs are hashed, so no original BSSIDs are visible

3 Discussion

3.1 Suitable machine learning algorithm

- What do we want to predict?
- We want to use supervised learning
- We want to predict the 3 best bssids for a given location of a user
- We could use regression or classification
- Pro regression: We could predict the exact signal strength
- Con: bssids could not be compared and though predicted
- Pro classification: bssids could be compared and though predicted

4 Implementation

4.1 Preprocessing

- Preprocessing important step of ML
- Data is not consistent
  - Wi-Fi and waypoint Data are not measured at the same time (some could be event triggered, but just speculation)
  - First: Interpolate waypoints to the timestamps of Wi-Fi data
  - Second: Merge interpolated waypoints and Wi-Fi Data
  - Detected jumps in time and in position
  - Present solutions: Split data into several parts, where the position more than 10 meters from the last position or time difference more than 60 minutes
  - time difference of more than 60 minutes could also lead to a jump in position: Therefore, position more than 10 meters away
  -

4.2 Machine Learning Model

5 Evaluation

5.1 Adapting parameters

- ...

6 Conclusion

6.1 Conclusion

- ...

Bibliography

[1] "IEEE Standard for Information technology– Local and metropolitan area networks– Specific requirements– Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 2: Fast Basic Service Set (BSS) Transition". In: *IEEE Std 802.11r-2008 (Amendment to IEEE Std 802.11-2007 as amended by IEEE Std 802.11k-2008)* (July 2008). Conference Name: IEEE Std 802.11r-2008 (Amendment to IEEE Std 802.11-2007 as amended by IEEE Std 802.11k-2008), pages 1–126. DOI: 10.1109/IEEESTD.2008.4573292.

[2] "IEEE Standard for Information technology– Local and metropolitan area networks– Specific requirements– Part 11: Wireless LAN Medium Access Control (MAC)and Physical Layer (PHY) Specifications Amendment 1: Radio Resource Measurement of Wireless LANs". In: *IEEE Std 802.11k-2008 (Amendment to IEEE Std 802.11-2007)* (June 2008). Conference Name: IEEE Std 802.11k-2008 (Amendment to IEEE Std 802.11-2007), pages 1–244. DOI: 10.1109/IEEESTD.2008.4544755.

[3] *Indoor Location & Navigation | Kaggle*. https://www.kaggle.com/competitions/indoor-location-navigation. (Visited on July 11, 2023).

[4] *Indoor Navigation: Complete Data Understanding*. https://kaggle.com/code/andradaolteanu/indoor-navigation-complete-data-understanding. (Visited on Apr. 25, 2023).

[5] *Kaggle: Your Home for Data Science*. https://www.kaggle.com/. (Visited on July 23, 2023).

Acronyms

Zusammenfassung

...

Eidesstattliche Erklärung

Hiermit versichere ich, dass meine Bachelor's thesis "Machine Learning-based User Movement Prediction in Layer 2 Networks" ("Vorhersage von Benutzerbewegungen in Layer 2 Netzwerken basierend auf Maschinellem Lernen") selbstständig verfasst wurde und dass keine anderen Quellen und Hilfsmittel als die angegebenen benutzt wurden. Diese Aussage trifft auch für alle Implementierungen und Dokumentationen im Rahmen dieses Projektes zu.

Potsdam, den 29. Juli 2023,

_____

(Lina Wilske)