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

August 3, 2023

Abstract

...

Contents

1 Introduction

In large-scale Wireless Fidelity (Wi-Fi) environments such as office buildings, shopping malls, and airports, where multiple Access Points (APs) are required, people often move around indoors with their mobile devices. To maintain a stable connection to the Service Set Identifier (SSID) a device is connected to, it must remain in range of the AP or may roam to another AP with the same SSID. Roaming has been an essential feature of Wi-Fi since the advent of the 802.11k[3] feature. This process was further enhanced with AP-initiated roaming, introduced in 802.11r[2]. However, the current roaming process doesn't account for human movement patterns. For example, if a station is moving away from AP1 towards AP2, ideally, AP2 should initiate the roaming process. Unfortunately, this is not feasible under the current scheme as AP1 cannot detect the station'a movement to AP2.

2 Background

Machine Learning (ML), a field of computer science known for predicting future events based on past ones, can potentially bridge this gap. Arthur Samuel was the first to use ML in 1959 to enable a computer program to improve its performance through self-play and learning from past decisions. [9]. Today, ML applications extend to diverse fields like image and speech recognition, vehicular networks[8], and human movement prediciton[1]. As Szott et al. noted in their survey[10], ML is now utilized in Wireless Local Area Networks (WLANs). For Wi-Fi, a prediction for a possible roam could initiate the roaming process sooner, thereby improving the user experience and overall Wi-Fi network performance.

ML models, however, require data to function. There are two primary data sources: generate new data or utilize existing data. Data generation necessitates a comprehensive plan accounting for technologies, interferences, and data setup and collection—this can be a time-consuming process. Consequently, this study will utilize pre-existing data from a 2021 competition by Microsoft Research[4]. The data will be analyzed to determine which segments are required for the ML model. Preprocessing is also necessary to convert raw sensor data into a format that the ML model can use.

Time series data, comprising a sequence of data points ordered in time, represents a common structure in many domains, including user mobility within a Wi-Fi network. Owing to its inherent temporal dependencies—where subsequent data points can be influenced by previous ones—particular machine learning techniques are typically employed. These include the Autoregressive Integrated Moving Average (ARIMA) model, and Recurrent Neural Network (RNN) models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). Each of these techniques is designed to capture and leverage temporal patterns within the data, facilitating the prediction of future trends based on historical observations.

In the context of machine learning model development, the configuration of hyperparameters represents a crucial task. Defined as the set of parameters that govern the learning process and are not learned from the data, hyperparameters encompass elements such as the learning rate, the number of layers within a neural network, and the quantity of clusters in a clustering algorithm. As these parameters are determined a priori, their careful selection—known as hyperparameter tuning or optimization—is necessary to maximize model performance. This iterative procedure involves exploring various hyperparameter combinations in search of the configuration that yields the most accurate predictions.

Moreover, the evaluation of data stands as an essential component of any machine learning project, and this remains true for time series prediction. Such evaluation involves an assessment of the model's performance by contrasting predicted outcomes with actual results. One common technique is cross-validation, wherein the data set is partitioned into a training set for model training, and a validation set for model evaluation. Performance

evaluation is indispensable for understanding the model's accuracy, reliability, and its ability to generalize to new, unseen data. Furthermore, it provides insights into potential underfitting, where the model fails to learn sufficient patterns from the training data, or overfitting, where the model becomes overly sensitive to noise or outliers in the training data, both of which can significantly impair predictive performance.

3 Data analysis

The dataset used in this thesis is the Indoor Location & Navigation from kaggle[7] which was part of a competition of Microsoft Research in 2021[4]. The data was recorded in shopping malls by 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. In the following the dataset and data will be analyzed.

3.1 Components of the dataset

As noted in the kaggle notebook "Indoor Navigation: Complete Data Understanding" [6] 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

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.

For this thesis, the submission files as well as the test folder will not be used, because our goal is another type of prediction.

3.2 File structure

Each file in each floor folder is a **.txt** file. The first two lines and the last are denoted with "#". The first contains the start time of the recording, the second site information SiteID as hash, SiteName, FloorId as hash and FloorName. The last line contains the end time of the recording. The main part of the data consists of the collected data. Each line contains a UNIX timestamp in milliseconds, followed by a data type and the data itself, which are all separated by a tab. Regard to the GitHub repository[5], the data type in the second column followed by its data can be one of the following:

(1) TYPE_ACCELEROMETER with x, y and z acceleration and an accuracy value

(2) TYPE_MAGNETIC_FIELD with x, y and z magnetic field and an accuracy value

(3) TYPE_GYROSCOPE with x, y and z gyroscope and an accuracy value

(4) TYPE_ROTATION_VECTOR with x, y and z rotation vector and an accuracy value

(5) TYPE_MAGNETIC_FIELD_UNCALIBRATED with x, y and z magnetic field and an accuracy value

(6) TYPE_GYROSCOPE_UNCALIBRATED with x, y and z gyroscope and an accuracy value

(7) TYPE_ACCELEROMETER_UNCALIBRATED with x, y and z acceleration and an accuracy value

(8) TYPE_WIFI with SSID, Basic Service Set Identifier (BSSID), Received Signal Strength Indication (RSSI), frequency, and last seen timestamp of the access point. The SSID and BSSID are hashed.

(9) TYPE_BEACON with Universally Unique Identifier (UUID), Major Identifier (MajorID), Minor Identifier (MinorID), Transmission Power (TxPower), RSSI, distance to the device measured by the beacon, Media Access Control (MAC) address and a timestamp as padding data. The MajorID and MinorID are hashed.

(10) TYPE_WAYPOINT with x and y coordinates which are the ground truth location labeled by the surveyor

Listing 3.1: A snippet of a file from the dataset

```
#   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
```

Each site has different amount of floors and also the amount of generated files varies. Each file contains different amount of waypoints and sensor data. The first and last data type in each file is a (8). Lines with types from (1) to (7) occur every 20 ms. *TYPE_WIFI*(8) occurs about every 1800 to 2050 ms.

As seen in Listing 3.1, the data are measured separately from each other, so there are no combinations of the data types. For machine learning it is essential to have all the data for a specific time for a time series. Especially, *TYPE_WAYPOINT* and *TYPE_WIFI* should be combined to get a location for the Wi-Fi data point. For a *TYPE_WIFI* data point a location is not provided. Furthermore, the *TYPE_WAYPOINT* data points are not evenly distributed as well as there are only 6549 of them in the floor folder with the most files, which is floor 1 of site 银泰城(城西店) which was hashed as "5d27075f03f801723c2e360f". A movement between these two data points may have occurred. For a connection between a *TYPE_WAYPOINT* and a *TYPE_WIFI* data point, an interpolation of waypoints between these data points could be useful. Therefore, this interpolation was made on the data resulting in Using linear interpolation of x and y coordinates of the waypoints for the timestamps of *TYPE_WIFI*.

This enables a more detailed analysis of the gathered data.

3.3 Visualization of one complete floor

- visualization of waypoints of one floor without interpolation
  - visualization of waypoints of one floor with interpolation

3.4 Peculiarities of the data

Further analysis of the dataset revealed some peculiarities, which are described in the following.

The data is collected by different devices, at different timestamps and days, and one problem for the ML could be, that the waypoint data were measured irregular.

1. Timestamp 1: 1572951273212.0, Timestamp 2: 1574853182904.0, Time Difference in ms: 1901909692.0 in min: 31698.494866666668
2. Timestamp 1: 1571745701735.0, Timestamp 2: 1572937550526.0, Time Difference in ms: 1191848791.0 in min: 19864.146516666668
3. Timestamp 1: 1571398890607.0, Timestamp 2: 1571745386405.0, Time Difference in ms: 346495798.0 in min : 5774.929966666667
4. Timestamp 1: 1571222697051.0, Timestamp 2: 1571393792128.0, Time Difference in ms: 171095077.0 in min : 2851.5846166666665
5. Timestamp 1: 1572947759794.0, Timestamp 2: 1572950695861.0, Time Difference in ms: 2936067.0 in min: 48.93445
6. Timestamp 1: 1572942316535.0, Timestamp 2: 1572944901430.0, Time Difference in ms: 2584895.0 in min: 43.081583333333334
7. Timestamp 1: 1572946047608.0, Timestamp 2: 1572947705343.0, Time Difference in ms: 1657735.0 in min: 27.628916666666665
8. Timestamp 1: 1572938081167.0, Timestamp 2: 1572938843518.0, Time Difference in ms: 762351.0 in min: 12.70585
9. Timestamp 1: 1571397963778.0, Timestamp 2: 1571398570240.0, Time Difference in ms: 606462.0 in min: 10.1077
10. Timestamp 1: 1572937620091.0, Timestamp 2: 1572938025654.0, Time Difference in ms: 405563.0 in min: 6.759383333333333

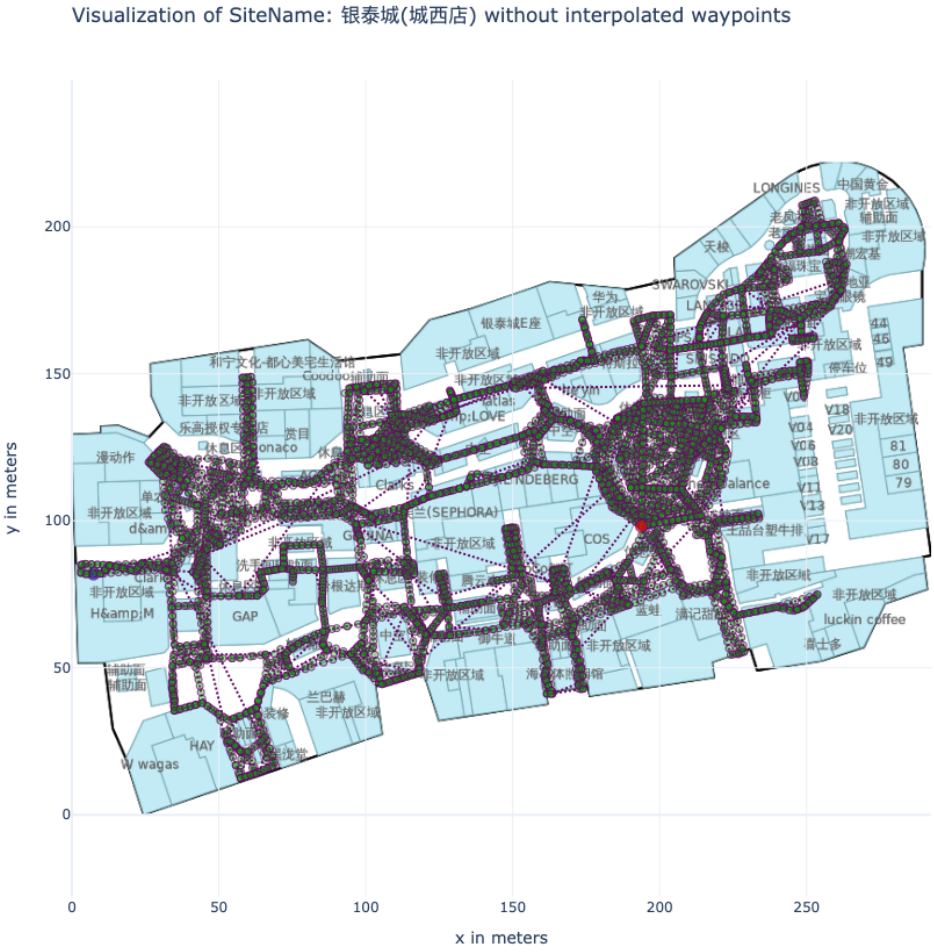

Figure 3.1: Visualization of the waypoints for SiteName: 银泰城(城西店)

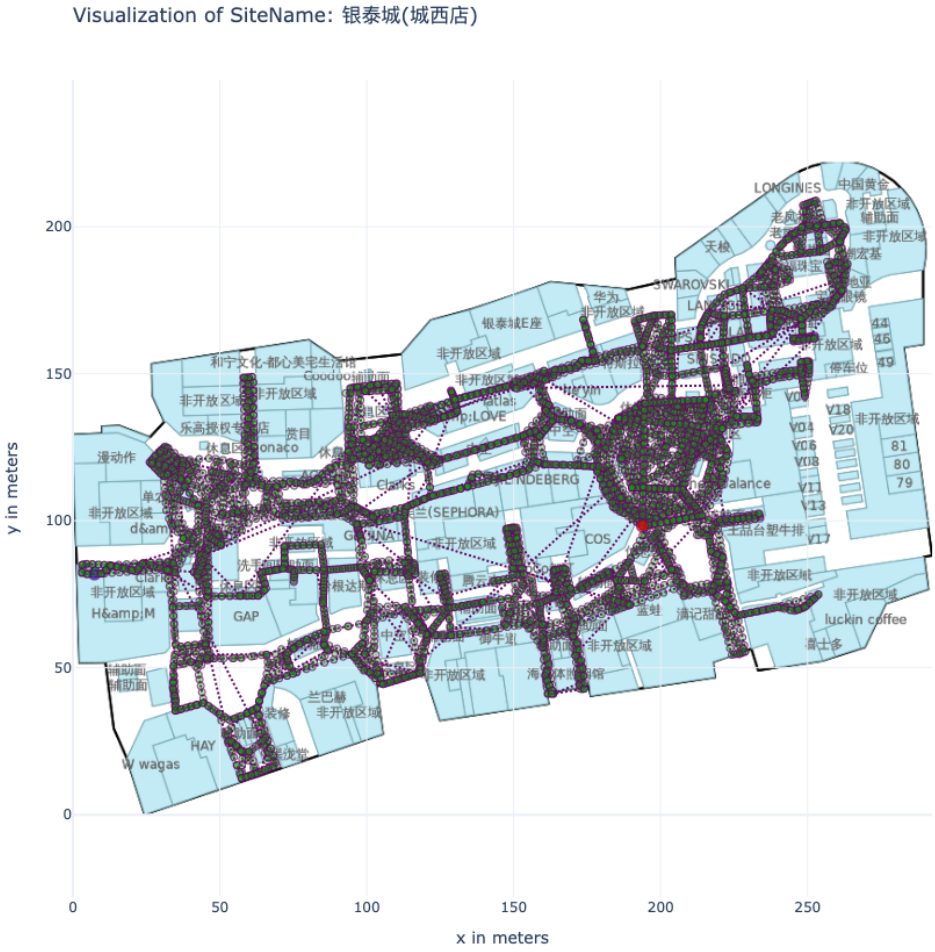

Figure 3.2: Visualization of the interpolated waypoints for SiteName: 银泰城(城西店)

4 Suitable Machine Learning Algorithm

As seen in chapter 3, we

4.1 X

4.2 Y

4.3 Z

5 Implementation

5.1 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
  -

5.2 Machine Learning Model

6 Evaluation

6.1 Adapting parameters

- ...

7 Conclusion

7.1 Conclusion

- ...

Bibliography

[1] Akinori Asahara, Kishiko Maruyama, Akiko Sato, and Kouichi Seto. "Pedestrian-movement prediction based on mixed Markov-chain model". In: *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. GIS '11. New York, NY, USA: Association for Computing Machinery, Nov. 2011, pages 25–33. ISBN: 978-1-4503-1031-4. DOI: 10.1145/2093973.2093979. URL: https://dl.acm.org/doi/10.1145/2093973.2093979 (visited on Apr. 26, 2023).

[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 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.

[3] "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.

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

[5] *indoor-location-navigation-20*. URL: https://github.com/location-competition/indoor-location-competition-20 (visited on July 31, 2023).

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

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

[8] *Machine learning in vehicular networking: An overview | Elsevier Enhanced Reader*. en. DOI: 10.1016/j.dcan.2021.10.007. URL: https://reader.elsevier.com/reader/sd/pii/S2352864821000870?token=8CB54BF40B6550C82C35044053EA5D7A34068CF3FC56DDC84C1D2C7C20854EC91213844D1E17CE0813A945E0642DC345&originRegion=eu-west-1&originCreation=20230426200105 (visited on Apr. 27, 2023).

[9] A. L. Samuel. "Some Studies in Machine Learning Using the Game of Checkers". In: *IBM Journal of Research and Development* 3.3 (1959), pages 210–229. DOI: 10.1147/rd.33.0210.

[10] Szymon Szott, Katarzyna Kosek-Szott, Piotr Gawłowicz, Jorge Torres Gómez, Boris Bellalta, Anatolij Zubow, and Falko Dressler. "Wi-Fi Meets ML: A Survey on Improving IEEE 802.11 Performance With Machine Learning". In: *IEEE Communications Surveys & Tutorials* 24.3 (2022). Conference Name: IEEE Communications Surveys & Tutorials, pages 1843–1893. ISSN: 1553-877X. DOI: 10.1109/COMST.2022.3179242.

Acronyms

AP	Access Point
ARIMA	Autoregressive Integrated Moving Average
BSSID	Basic Service Set Identifier
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
MAC	Media Access Control
MajorID	Major Identifier
MinorID	Minor Identifier
ML	Machine Learning
RNN	Recurrent Neural Network
RSSI	Received Signal Strength Indication
SSID	Service Set Identifier
TxPower	Transmission Power
UUID	Universally Unique Identifier
Wi-Fi	Wireless Fidelity
WLAN	Wireless Local Area Network

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 3. August 2023,

_____

(Lina Wilske)