$P_{bMFS}$ - PERIODICITY BASED MOBILITY FORECASTING SYSTEM

A Dissertation in
Computer Science and Engineering
by
Bhaskar Prabhala

© 2015 Bhaskar Prabhala

Submitted in Partial Fulfillment
of the Requirements
for the Degree of
Doctor of Philosophy

December 2015
The dissertation of Bhaskar Prabhala was read and approved\(^1\) by the following:

Thomas F. La Porta  
Professor of Computer Science and Engineering  
Dissertation Adviser  
Chair of Committee

Raj Acharya  
Professor of Computer Science and Engineering  
Head of the Department of Computer Science and Engineering

Wang-Chien Lee  
Associate Professor of Computer Science and Engineering

Soundar Kumara  
Professor of Industrial and Manufacturing Engineering

\(^1\)Signatures on file in the Graduate School.
Abstract

We present synthesized findings from a systematic study of user mobility based on well grounded data sets. We leverage human periodicity and the practice of mining attributes of place-to-place transitions. Next place predictions are the atomic units in constructing end-to-end user mobility trajectories based on historical trace data. We start with a baseline of the user’s current place, start time, and end time to predict the next place. We demonstrate the efficiency of our algorithms, PeriodicaB and PeriodicaS, through aggregated average prediction accuracies across all users over a large set of diverse participants. PeriodicaS mines periodicity intelligently in users’ mobility traces and further improves prediction accuracies with additional classification rules. We derive these classification rules by applying explicit semantic annotations (home, work place and public transportation points associated with places visited), and accompanying group information. We propose novel ways of transforming bits of information in the mobility traces, defined as inherent semantic annotations, as features for mobility modeling in PeriodicaS. Inherent semantic annotations are deduced from temporal variations in visited places such as end time only, and measuring duration time. We deduce more inherent semantic annotations from place rankings by frequency of visits. By progressively employing the semantic annotations, explicitly stated in the data set and deduced from the mobility traces, we improve next place prediction accuracies up to 58% compared to baseline predictions. We cluster the users’ visited places based on their proximity from each other make next place predictions based on these clusters. We make multiple choice for next place predictions and show, while considering three choices for each next place prediction, the aggregated average prediction accuracies reach 94.36% accuracy compared to baseline accuracy average of 50.47%.

PbMFS - Periodicity based Mobility Forecasting System - is built around our next place prediction algorithm PeriodicaS. PbMFS generates end-to-end trajectories of user mobility and forecasts users’ movements involving places of interest. Deduction and effective usage of inherent semantics is the distinguishing characteristic of our forecasting system. While forecasting a user’s mobility trajectory, PbMFS dynamically chooses the appropriate feature set for a given next place prediction and switches between feature sets during the sequence of predictions as required. Forecasts generated by PbMFS lead to realistic models of opportunistic networks for service deployment and contingent strategies for redeployment.
# Table of Contents

List of Tables ........................................ vi
List of Figures ........................................ vii
Acknowledgments ....................................... ix

Chapter 1. Introduction .................................. 1
  1.1 Thesis Statement .................................... 2
  1.2 Contributions ....................................... 3
  1.3 Overview ........................................... 4

Chapter 2. Concepts, Methodology, Algorithms, and Data Sets .............. 6
  2.1 Periodicity .......................................... 7
  2.2 Processes and Techniques ......................... 10
    2.2.1 Period Detection in Trace through Overlay .... 10
    2.2.2 Prediction Accuracies ......................... 10
  2.3 Algorithms .......................................... 10
    2.3.1 Multi-Class Classification ..................... 11
    2.3.2 Periodicity Based Algorithm: PeriodicaB .... 12
  2.4 Data Sets ........................................... 14
    2.4.1 Nokia MDC Data Set ............................ 14
    2.4.2 Wireless Topology Discovery Data Set ......... 14
    2.5 Mobile Data Challenge by Nokia ................. 14

Chapter 3. Next Place Predictions ................................ 17
  3.1 Introduction ........................................ 17
  3.2 Related Work ...................................... 18
  3.3 Data Analytics ..................................... 19
  3.4 Predictions and Analysis ......................... 21
    3.4.1 Accuracy Profiles .............................. 23
    3.4.2 Analysis ....................................... 23
  3.5 Discussion ......................................... 27

Chapter 4. Semantic Annotations ................................ 28
  4.1 Introduction ....................................... 28
  4.2 Related Work ...................................... 29
  4.3 PeriodicaS and Feature Sets ...................... 30
    4.3.1 Preprocessing Mobility Traces ................ 32
  4.4 Collected Semantic Information .................... 32
    4.4.1 Accuracies with Place Labels and Companion Information 33
    4.4.2 Demographics .................................. 33
4.5 Inherent Semantic Annotations ........................................... 35
  4.5.1 Temporal Variants .................................................. 35
  4.5.2 Place Ranking ...................................................... 36
4.6 Summary and Discussion .................................................. 37

Chapter 5. Mobility Forecasting .............................................. 42
  5.1 Introduction .......................................................... 42
  5.2 Related Work and Background ....................................... 43
  5.3 Place-to-place Transitions and Annotations ......................... 45
    5.3.1 Visualization ................................................... 45
    5.3.2 Existing Annotations ........................................... 45
    5.3.3 Inherent Semantic Annotations ................................ 47
    5.3.4 Most Frequently Visited Places ................................ 49
  5.4 Inherent Semantic Annotations based on Distance Matrices ........ 51
    5.4.1 Clustering Visited Places ...................................... 51
    5.4.2 Results .......................................................... 53
  5.5 Multiple Choices for Next Place Predictions ....................... 54
    5.5.1 First, Second, and Third Choice Predictions for Next Place ... 56
    5.5.2 Prediction Accuracy Averages .................................. 57
  5.6 Individual User Prediction Accuracies ............................... 59
    5.6.1 User Participation ............................................... 59
    5.6.2 Selected Feature Sets and Specific Users ..................... 61
    5.6.3 Prediction Accuracies with Temporal and Spatial Annotations 62
  5.7 Mobility Forecasting ................................................. 64
    5.7.1 Operational Flow of Pn MFS .................................... 64
    5.7.2 End-to-end trajectories ........................................ 65
    5.7.3 Forecasts for Places of Interest ................................ 65
    5.7.4 Trajectories with Multiple Choices for Next Place Predictions 68
  5.8 Summary and Discussion ............................................. 71

Chapter 6. Models for Opportunistic Networks and Moving Forward ........ 72
  6.1 Opportunistic Networks ............................................ 72
  6.2 Summary and Conclusions .......................................... 73
  6.3 Future Work ........................................................ 74

Bibliography ........................................................................ 75
List of Tables

3.1 Aggregated prediction average accuracies across all users in the MDC and the WTD data sets using the next place predictors PeriodicaB and PeriodicaS .......................................................... 21

4.1 Place Labels and Descriptions ................................................. 30
4.2 Prediction Accuracies with Temporal Variants ....................... 35
4.3 Place Ranking and Prediction Averages ............................... 37

5.1 Prediction Accuracies Utilizing Existing Annotations from the Data Sets ........................................................... 47
5.2 Prediction Accuracies Utilizing Inherent Semantic from the Trace Data ............................................................... 48
5.3 Relative Spatial Distances between Places ......................... 51
5.4 Aggregated Prediction Accuracy Averages Clustering Visited Places Based on Distance Matrices .................................... 53
5.5 Aggregated Averages with Multiple Predictions ................. 57
5.6 Statistics from Participants’ Mobility Traces ....................... 59
5.7 Comparing Accuracy Averages of All Users, User #47, and User #51 with 15 Most Frequently Visited Places and Multiple Choices for Next Place Predictions ........................................... 61
List of Figures

2.1 Atomic Place-to-place Transition ........................................... 6
2.2 Sample Mobility Graph with User Movements .............................. 9
2.3 Intuition of the Periodicity Model from [29] ............................... 13
2.4 Visit Frequency Distribution over Places ................................ 13
2.5 Structure and Tables of MDC Data Set from [30] ......................... 15
2.6 Map of the campus with location in WTD Data Set from [27] ......... 16

3.1 Frequency distribution of the next place visited after place 3 and place 13 20
3.2 Accuracy of predictions for users in MDC data set ......................... 22
3.3 Accuracy of predictions for users in WTD data set ......................... 24
3.4 Accuracy of predictions VS amount of trace data for each user in MDC data set ........................................ 25
3.5 Error Analysis for User 8 ....................................................... 26

4.1 Baseline Prediction Accuracies for all Users ................................. 31
4.2 Aggregated average accuracies (as compared to the baseline of 50.47%) 34
4.3 Prediction accuracies for four different groups; baseline average is 50.47% 39
4.4 Number of unique place IDs visited by users ............................... 40
4.5 Distribution of top 10 places visited .......................................... 41

5.1 Mobility Traces of user #39 with all 128 places ........................... 46
5.2 Mobility Traces of User #39 with the Top Ten Frequently Visited Places and the Rest of the Places Combined into an Incidental-place (11) ... 48
5.3 Prediction Accuracies with Two Most Dominant Places - Average Accuracy is 71.35% as Compared to the Baseline Average of 50.47% .... 50
5.4 Prediction Accuracies for the Most Dominant Place - Average Accuracy is 81.77% as Compared to the Baseline Average of 50.47% .......... 50
5.5 Aggregated Average Accuracies for all Users with Sixteen Clusters Using Greedy Algorithm; Aggregated Average Accuracy Across all Users is 65.36% ...................................................... 55
5.6 Aggregated Averages for all Users with Sixteen Clusters Using Hierarchical Clustering Algorithm; Aggregated Average Accuracy Across all Users is 79.47% ...................................................... 55
5.7 User Average Accuracies with Three Predictions for Each of the Next Place Prediction - 15 Most Dominant Places; All Prediction Ave 13.61%, First Choice Prediction Average 50.85%, Second Choice Prediction Average 73.30%, Third Choice Prediction Average 8.83% .............................. 58
5.8 User Average Accuracies with Three Predictions for Each of the Next Place Prediction - Using Hierarchical Clustering Algorithm All Prediction Average 94.36% First Choice Prediction Average 79.47%, Second Choice Prediction Average 11.55%, Third Choice Prediction Average 3.02% ......... 60

5.9 Individual Prediction Averages across Five Different Feature Sets ......... 63

5.10 Individual User Prediction Accuracy Averages ................. 63

5.11 End-to-end Trajectory Example for User #14 .................. 66

5.12 End-to-end Trajectory Example for User #26 with Switching of Feature Vectors at A and B ............... 66

5.13 Next Place Predictions Choices with Current Place, Target Place, and Incidental Place ................. 66

5.14 Forecast for User #58 with Three Choices for Next Place Predictions at Each Hop; Next Place Predictions Matching with the Ground Truth are Shown in Gray .................. 67

5.15 Forecast for User #66 with Three Choices for Next Place Predictions at Each Hop; Next Place Predictions Matching with the Ground Truth are Shown in Gray .................. 69

6.1 User’s opportunistic network connections through the day: at home, work, lunch and other visited places ............. 70
Acknowledgments

This dissertation is a product of years of careful and methodical thought verified through empirical experimentation. I received much assistance during my journey.

I appreciate the support from my committee members Professor Raj Acharya, Professor Wang-Chien Lee and Professor Soundar Kumara. Special thanks to Professor Acharya for encouraging me to return to Penn State and finish the job that I started 39 years ago.

I am grateful to my adviser, Professor Thomas La Porta, for placing trust in me, enabling me to become a researcher, and guiding me throughout the journey. At every turn, Professor La Porta was there pointing me in the right direction and guiding me through difficult times both academically and personally. I learned much from Professor La Porta and am indebted to him.

James Edwards, Srikar Tati and Stephen McLaughlin, fellow researchers from the Institute for Networking and Security Research, reviewed my drafts and made helpful suggestions. Scott Rager spent many hours helping me improve my technical writing and presentation. He was generous with his time and thoughtful in providing me with his feedback. Nicolas Papernot, Nan Hu, and Stefan Achleitner were always eager to lend a hand with many systems related issues.

My research was supported by the U.S. Army Research Laboratory under the Network Science Collaborative Technology Alliance, Agreement Number W911NF-09-2-0053. I acknowledge the support of the participating members of the alliance.
Dedication

To my parents, Prabhakara Sastry and Seshamma.

To Chinna Mamayya, Kowtha A. Murthy.

To Sasi.

We are travelers on a cosmic journey, stardust, swirling and dancing in the eddies and whirlpools of infinity. Life is eternal. We have stopped for a moment to encounter each other, to meet, to love, to share. This is a precious moment. It is a little parenthesis in eternity.

Paulo Coelho, The Alchemist
Chapter 1

Introduction

As stated in a report by the National Research Council [1], Network Science is the study of network representations of physical, biological, and social phenomena leading to predictive models. Network Science is taking shape as an interdisciplinary field that explores many variations of complex networks. Several fundamental concepts are brought to bear on investigative efforts in Network Science dealing with networks that evolve over time and that are intrinsically complex in their infrastructure and behavior. During the last two decades, scientific research has produced many collaborative studies in this field [2].

Topics of network theory, interconnectivity and related issues, and the impact of network science are covered extensively in a recent book [3], which illustrates how networks are at the heart of complex systems. Developments in the field of analyzing the structure and function of complex networks are reviewed in [4]. Network analysis is a powerful way of studying the structure, representation and organization of a network, as shown in [5], and to recognize and interpret underlaying patterns; as part of Network Science these techniques are applied in phenomena as diverse as social interactions, neuron connections and the Internet. The first step for such an analysis is choosing the right network representation for the problem being investigated.

Mobility is an universal property, and many forms of mobility are studied in different disciplines of scientific research. The universe started with a big bang and has been in perpetual motion ever since. Galaxies, including our own, continue to move away from each other at ever increasing speeds. Recent research [6] suggests that the motion of the Milky Way will result in a collision with the Andromeda galaxy, forever altering the shape and nature of our galaxy. The impending collision and its aftereffects on our own Sun and the Solar System are described in great detail in [7].

Humans have also been shaped by our own mobility, starting with our journey out of Africa and then by moving across the globe. Mobility of early modern humans, as well as many other factors, contributed to our present state of complexion [8]. Humans’ capability and desire for mobility continues to be studied at various levels of granularity and impacts both our present and our future.

With the advent of smart mobile devices, we are able to capture individual mobility on a minute-by-minute basis. Under the umbrella of location services, smart devices capture location and time associated with users’ mobility. Many other bits of temporal and spatial information directly or indirectly related to users’ place-to-place transitions are constantly collected by their smart devices.

Mobility traces are represented as a network with nodes corresponding to place-IDs and edges representing the corresponding place-to-place transitions. Place-to-place transitions for individuals are collected in a data set, and place coordinates are typically
anonymized for privacy concerns. User mobility networks thus constructed from a few months of place-to-place transitions and temporal values (start time, end time, and duration times at each place-id) become complex. These networks contain patterns of user movements, which when extracted provide models for user-specific mobility.

We start with raw mobility traces, then design a next place predictor, next define and deduce inherent semantics, and finally demonstrate a serviceable mobility forecasting system. We study user mobility patterns with two overarching goals:

- Develop realistic and useful models for opportunistic networks, a type of mobile ad hoc networks, for service deployment and redeployment strategies.
- Anticipate accurately individual users’ future mobility to effectively guide, recommend, and provide required intelligence.

Anticipating users’ future actions or mobility is essential in providing beneficial communication service. Given the unpredictable nature and behavioral patterns of human beings, predicting their needs and actions with reasonable accuracy is a challenging task. However we find foundational periodic patterns and rhythmic cycles in human mobility. Our ability to synthesize and deduce inherent semantic annotations from the captured traces greatly assists us in improving users’ next place predictions. Our PbMFS - Periodicity based Mobility Forecasting System - exploits these periodic patterns and the synthesized semantic annotations through available data mining techniques, and forecasts users’ future mobility consistently and reliably. PbMFS generates a backup option for the portions of the forecast where a clear trend can not be extrapolated from the captured traces.

1.1 Thesis Statement

Accurately predicting a person’s next place destination, with no consideration for place semantic annotations, is a vexing problem; several techniques and algorithms as point solutions exist in the literature. We efficiently recognize and interpret periodic patterns in user mobility traces. Through effective usage of data mining techniques and semantic annotations from user mobility traces, we generate consistently reliable next place predictions.

Performance of next place predictors can be improved by mining relevant information associated with user’s mobility traces. We define and deduce inherent semantic annotations, which are more reliable than user-supplied place semantic information, from these place-to-place transition. Our next place predictor, PeriodicaS, generates significantly improved next place predictions. We demonstrate the effectiveness of our next place prediction algorithm through aggregated prediction accuracy averages over a diverse group of users with mobility traces collected over a period of 12 months.

Proximity between users’ visited places is a powerful semantic annotation; we deduce this information from the distance matrices from the data set. Clustering the visited places based on their proximity improves the next place prediction accuracies significantly.
PeriodicaS uses the multi-class classification scheme of Support Vector Machines for making next place predictions. We extend PeriodicaS to make second choice next place predictions and give better accuracy when both the choices are compared with the ground truth.

PeriodicaS, with the cumulative benefit of the above mentioned techniques, approaches the accuracy of 93% in predicting users’ next place destinations. Song et al [2] established 93% predictability, a formidable target, as the performance barometer for next place predictors based on users’ periodicity through Song et al’s experiments and framework.

With PeriodicaS as the core component, we built PbMFS, Periodicity based Mobility Forecasting System. We constructed an evolving database of feature sets based on existing and deduced semantic annotations, and the corresponding individual user’s next place prediction average accuracies. PbMFS is a versatile and flexible system that adopts and switches between the best suited feature sets appropriately while generating a forecast for an individual user.

Publications Related to This Work:

- Bhaskar Prabhala, Jingjing Wang, Budhaditya Deb, Thomas La Porta, and Jiawei Han. Leveraging periodicity in human mobility for next place prediction. In Wireless Communications and Networking Conference (WCNC), 2014 IEEE, pages 2665 – 2670, April 2014.
- Bhaskar Prabhala and Thomas La Porta. Next Place Predictions Based on User Mobility Traces. Computer Communications Workshops (INFOCOM WKSHPS), 2015 IEEE Conference.
- Bhaskar Prabhala and Thomas La Porta. Mobility Forecasting for Generating Models of User Opportunistic Networks. Under preparation for a journal submission.

1.2 Contributions

- From the commonly used steps of Knowledge Discovery in Databases (KDD), and data mining techniques, we leverage periodicity in user movements to generate models and make next place predictions.
- We consolidate these techniques, described above, into two algorithms, PeriodicaB and PeriodicaS, and analyze the resulting predictions on two data bases with place-to-place transitions collected over a long period (four to nine months) from a diverse group of participants.
• We consider available place semantics and demographic information for improving the accuracy of next place predictions. We deduce inherent semantic annotations through data analytics and enhance our feature sets with these annotations. We show prediction accuracy gains from incorporating these semantic deductions in PeriodicaS.

• We generate fifteen to twenty feature sets from available place semantic information in the trace data sets and deduce semantic annotations, and we analyze the aggregated average prediction accuracies corresponding to these feature sets. In addition, we examine individual user next place prediction average accuracy, which is needed for the forecasting system. We introduce the concepts of clustering visited places based on their proximity from each other and multiple next place predictions and analyze the aggregated average prediction accuracies.

• We developed PbMFS - Periodicity based Mobility Forecasting System - with our next place predictor, PeriodicaS, and the database of the feature sets as the core components. We describe the operational flow of PbMFS through generating users’ mobility forecasts: end-to-end trajectories, patterns from specific places of interest, and trajectories with backup next place predictions. We compare the accuracy of the generated forecasts from ground truth (previously unseen by our next place predictor) from the users’ traces.

1.3 Overview

Chapter 2 presents the concept of periodicity, techniques from data mining that inspired our methods, and modeling and prediction algorithms. We discuss network representation of user mobility networks for our analysis and synthesis. We describe the Nokia MDC and the WTD data sets used by our algorithms. We share the experience of participating in a global Mobile Data Challenge, our starting point in this journey. We also provide a summary of other challenge tasks, and provide a summary from the associated workshop.

From our participation in the challenge, we consolidate our techniques into two next place prediction algorithms, PeriodicaB and PeriodicaS. We describe the results and analysis of next place predictions on two popular data sets of mobility traces, MDC and WTD data sets, in Chapter 3.

We introduce the construct of inherent semantic annotations in Chapter 4 and show how to deduce them from place-to-place transitions. We utilize the available place annotations and deduced semantic annotations in improving next place predictions. We show the improvements in prediction accuracies and analysis of these improvement in this chapter.

Chapter 5 presents a systematic approach to deducing relevant and reliable information associated with users’ traces through inherent semantic annotations. In addition, we introduce and demonstrate the efficiency of two new techniques for enhancing the accuracy of next place predictions:
• We deduce inherent semantic annotations based on distance matrices. We propose two algorithms, Greedy and Hierarchical Clustering, to segment users’ visited places based on their proximity to each other. Taking advantage of the results from these algorithms, PeriodicaS achieves an improvement of 58% in aggregated prediction accuracy averages across all users in the data set.

• We capture the intermediate calculations of LIBSVM [9] to make multiple choices for next place predictions. We designed a new algorithm, using PeriodicaS as the basis, for making 1st, 2nd, and 3rd choice predictions for users’ next place. When the three choices for next place prediction are taken into account while comparing them with the ground truth, we achieve a staggering 94% aggregated average prediction accuracy across all users in the data base.

In Chapter 5 we also introduce our mobility forecasting system, P_bMFS, based on our next place predictor. After describing the underlying components and the operational flow of P_bMFS, we take up a few examples of generated forecasts - end-to-end trajectories, specific place-to-place movements, and forecasts with backup trajectories - to demonstrate the system’s usability.

We summarize our findings and insights in Chapter 6. We detail future directions for our work with a) enhancements to our next place predictor, b) improvements to forecasting, and c) applicability of our system in emerging smart digital assistants. We propose a mechanism for generating models for users’ opportunistic networks that are useful for providing communication services such as store-carry-and-forward and service (re)deployment strategies.
There has been a flurry of research activity during the last ten years studying human mobility. Relationships, organizations and structure among people lend themselves naturally to representation as networks. Such networks are studied extensively for community detection, substructures, and other patterns [10]. Human mobility is studied either through generated mobility models or through individuals' mobility history.

Human mobility models are studied in [11] through generation of graphical models. Models for mobility networks are generated and characterized using well known models such as Random Walk [12], Random Waypoint [13], Reference Point Group Mobility (RPGM) [14] [15], and Manhattan [16]. Medina et al. propose the Universal Mobility Modeling Framework (UMMF) in [17] that generates specific mobility models based on user supplied parameters. With this frame, and based on Repeated Traversal, Bounding Overwatch, and Pincer Movement military strategies, tactical mobility models are generated in [18]. Using the models generated from these studies, data replication and other communication service algorithms are analyzed in [19]. Models generated through such efforts can be studied in order to provide a variety of services, but they portray sunny-side scenarios.

Mobility traces are collected by many groups tracking user movements through location services of their smart mobile devices. These collected traces are grouped into data sets and are available for research purposes. Sometimes these data sets contain information about place semantics and user actions associated with places. The site http://crawdad.cs.dartmouth.edu/index.html has an archive of many such trace data sets collected through wireless devices. LiveLabs is a test bed for experiments and next place predictions, and consists of mobility traces collected in an indoor campus environment [20] [21].

Place-to-place translations as shown in Figure 2.1 with no semantic annotations, form the basis for a mobility trace. Mobility traces typically contain noise and are incomplete, as some of the location information may be lost or the mobile device switched off through some place-to-place transitions.

Fig. 2.1. Atomic Place-to-place Transition
We start with mobility traces collected over a period of time, recognize existing patterns, exploit properties of periodicity, and deduce inherent semantic annotations to predict next place destinations. Using these predictions we build users’ end-to-end trajectories that form the base models for service-providing opportunities.

Learning from a user’s mobility traces, there are two approaches for anticipating the user’s location: a) predict the user’s location for a time-slot based on the distribution of time spent by the user at frequently visited places, and b) give the current time and the user’s current place and predict the user’s next destination based on place-to-place transitions. These two variations of predicting a user’s next place – using the current time only, versus given using the current time and current place – are explored and compared in [22]. Predictions based on both approaches have their uses in different applications.

Accurately and consistently predicting place-to-place transitions (the second approach described above) is much harder than making next place predictions considering the distribution of time spent by users at frequently visited places. By definition, predictions based on the second approach are required to be more precise. Our next place predictions are always based on place-to-place transitions with current place and time.

Difficulties with studying motion as articulated in The Unknowable of First Principles by Herbert Spenser [23] apply well to human mobility: “So with the divisibility of space and time; both of these are ultimately irrational ideas. Motion is wrapped in a triple obscurity, since it involves matter changing, in time, its position in space.”

Service providers need to anticipate users’ needs, desires and actions to improve their services to the users; understanding users’ mobility and anticipating their next place prediction(s) is an important step. The ability to anticipate users’ mobility improves the functionality of digital assistants, such as Siri and Cortana, significantly [24] [25].

Synthesizing and learning from human mobility traces is essential for effective next place predictions. The first step in that process is to develop mobility models from daily mobility history. In [26], Lee et. al. present mobility models that bring out the mobility patterns from human mobility history. Learning and modeling based on traces of place-to-place transitions over a period of time pose problems in recognizing mobility patterns, as parts of the trace can become irrelevant and generate false positive patterns based on user lifestyle changes. We discuss these issues in Section 2.5.1.

This chapter describes the underlying concepts, processes and techniques for our methods and modeling. We give a general description of PeriodicaB and PeriodicaS, our periodicity-based next place prediction algorithms. We expand on the usage of these algorithms and other implementation details further in the following chapters. We use the MDC data set from Nokia and the WTD data set from UCSD [27] for our studies. We describe these data sets in detail in Section 2.5.

### 2.1 Periodicity

Periodicity is defined as a quality, state, or fact recurring at regular intervals, such as every 24 hours. The concept of periodicity has been known for a long time, and in his seminal work [23], Will Durant elaborates on the rhythm of motion, while presenting Herbert Spencer’s [1820-1903] all-encompassing conception of evolution: “All
nature is rhythmical, from the pulsations of heat to the vibrations of violin strings; from the undulation of light, heat and sound to the tides of the sea; from the periodicities of sex to the periodicities of planets and comets and stars; from the alternation of night and day to the succession of the seasons, and perhaps to the rhythms of climactic change; from the oscillations of molecules to the rise and fall of nations and the birth and death of stars."

We are interested in periodic patterns that exist in human movements and are part of individuals’ daily mobility histories. In [2], Song et al. find that there is 93% potential predictability in user mobility. They studied two data sets, an anonymized data set representing 14 weeks of call patterns from 10 million mobile phone users, and a second set of anonymized records of 1000 users, whose coordinates were recorded every hour over eight days. Despite the inherent population heterogeneity in their study, they found that maximal predictability varies very little between a high of 93% and a low of 80%.

Through their examination of collected users’ smartphone data [28], Montoliu et al. discovered places of interest in their daily lives: users tend to stay at home 67% of their time, at a place related to their work or school 20% of the time, and at a place associated with free time 7% of their time each day. They spend 5% of their time at other kinds of places and at unknown places 1% of their time.

We generated the periodicity graph, shown in Figure 2.2, which is a composite representation of daily movements for a subset of participants in the MDC dataset. The solid lines represent the place-to-place transitions between the most frequently visited places and the dotted thin lines represent the edges leading to destinations that are visited once in a while.

From the trace we examined, users visited between 25 and 199 places over a period of 11 months. We took the top 10 frequently visited places for the periodicity graph in Figure 2.2. We identify the other kinds of places and the unknown places (where users spend a combined 6% of their time) and label them as "Inc." in the periodicity graph. The thickness of the edges represents the frequency of visits between the places.

We reexamine the two approaches of predicting a user’s next place with the periodicity graph in Figure 2.2. The percentages of times spent at home, at work etc., are computed from users’ corresponding time intervals in the mobility traces. From the query with respect to users location, matching the time-slot to a time-interval and the percentage of time spent, a user’s next place can be predicted through efficient look-ups and comparisons.

Every edge corresponding to a respective place-to-place transition in the periodicity graph in Figure 2.2 is a simplified representation of the associated temporal information from users’ mobility traces. We explore the visual representation of users’ raw mobility networks from traces. The temporal requirement contained in the query, essentially forcing the usage of the second approach, introduces further complexities in predicting users’ next place. In the rest of the thesis, we make next place predictions based on place-to-place transitions with current place and time, and analyze the resulting predictions and the averages.
Fig. 2.2. Sample Mobility Graph with User Movements
2.2 Processes and Techniques

In this section, we describe the techniques used in modeling the trace data. We detail the processes that form the basis for our algorithm development and methods used for analyzing the performance of our algorithms.

2.2.1 Period Detection in Trace through Overlay

There is periodicity in user movements as described in the previous section. Our challenge is in determining the detection of usable periods in these place-to-place transitions. First we present the period detection technique through an example and explain how we apply the technique to our algorithms.

The technique and the example are taken from [29]. In the example from their finding, reproduced in Figure 2.3, an event has a period 20 (the event happens between 20k+5 to 20k+10) and we have a sequence of eight observations of that event. If we overlay the observations with the correct period, we can see that the observations will form dense clusters along the timeline, which gives a time distribution of the periodic events.

We observe that daily, weekly and hourly periods are natural in user mobility traces. In addition, we use the periods of morning, afternoon, and evening as periods in classifying our place-to-place transitions. We take the start time and end time and map them into features or attributes based on hour of the day, morning, etc., for our algorithms.

2.2.2 Prediction Accuracies

We characterize users’ mobility patterns through a systematic study of their place-to-place transitions, and we always predict a user’s next place destination given the context of the current place and time. First, we compute aggregate prediction accuracy for one user by counting all the next place predictions with the ground truth. There is no partial credit for predicting a nearby place.

We aggregate individual prediction accuracies for all the participants in the data set to come up with an aggregated prediction average. We use this measure to characterize our ability to mine user periodicity in collected place-to-place transitions through the PeriodicaB and PeriodicaS algorithms.

2.3 Algorithms

Both PeriodicaS and PeriodicaB are designed to follow the basic steps in the Knowledge Discovery in Databases (KDD) process. We use a pre-processing step to convert each place-to-place transition to a set of attributes: start day of week, end day of week, start hour of day, end hour of day, start minute, end minute, current place id, normalized start time, and a user’s next place. These algorithms were initially formulated as part of our participation in the MDC competition by Jingjing Wang.

Since the data does not contain relationship information between users, we build user-specific models. Based on these models, PeriodicaS and PeriodicaB predict the
next place given a user's current context. We take the multi-class classification approach, building a feature vector for each place-to-place transition with next place as the label for Periodica\(S\), while Periodica\(B\) is built by mining periodicity in a user’s place-to-place transitions directly.

2.3.1 Multi-Class Classification

With all of the user’s current information, we model our approach into a multi-class classification problem. For each entry, we extract features from the current context and use the next place as the label. We highlight a few issues here. First, the class labels are highly imbalanced since there are dominant places. Meanwhile, many of the minority classes have very few data samples, e.g., only 1 or 2 samples. This will cause the classifier to favor the majority classes. And since there are not enough samples for the minority classes, the accuracy of the minority classes will be very low, as we will see later in the analysis.

We describe the operational flow of Periodica\(S\) in Algorithm 1 referring to the functional modules of our system. We take the user’s mobility traces and split them into a training sequence and a test sequence, and reserve a portion of the sequence as ground truth for comparison with the generated next place predictions. Our next place predictor never sees the third sequence of traces described above, Periodica\(S\) uses them only for comparing and computing prediction accuracy.

The training sequence, test sequence and the ground truth are supplied to Periodica\(S\) as input csv files or vectors: user − seq.csv, tes − seq.csv, and grnd − truth. Periodica\(S\) returns the user’s average next place prediction accuracy as acc − ave.

Algorithm 1: Periodica\(S\)

\[
\begin{align*}
\text{Input: } & \text{user} \rightarrow \text{seq}, \text{tes} \rightarrow \text{seq}, \text{grnd} \rightarrow \text{truth} \\
\text{Output: } & \text{acc} \rightarrow \text{ave} \\
1 & \text{acc} \rightarrow \text{ave} \leftarrow 0; \\
2 & \text{featureV, labelV} \leftarrow \text{map(user - seq - csv)}; \\
3 & \text{cross validate by dividing the training sequence into five partitions; } \\
4 & \text{for } i \leftarrow 1 \text{ to } 5 \text{ do} \\
5 & \quad \text{best} (\text{KernelParams}) \leftarrow +; \\
6 & \quad \text{model} \leftarrow \text{svmtrain(featureV, labelV, KernelParams)}; \\
7 & \text{for } i \leftarrow 1 \text{ to length(test - seq) do} \\
8 & \quad \text{if svmpredict(model, test - seq[i] == grnd - truth[i] then} \\
9 & \quad \quad \text{acc - ave} \leftarrow +; \\
10 & \text{return acc - ave;}
\end{align*}
\]

First, we translate the training sequence into a feature set and corresponding label; these routines base the translations on the place-to-place transitions in the trace lines
of the csv files, utilize the place semantic information and deduces inherent semantics appropriately. Through line five, we invoke the function that computes the $coeff0$ and $gamma$ values (parameters to the RBF Kernel employed by LIBSVM). We accumulate the best parameters while validating the predictions for partition $i$ after training with the other four partitions.

Secondly in line six, we build the required model for making next place predictions using the svmtrain function of LIBSVM. We use the best $RBFKernelParameters$ computed during the cross validation step.

Finally in lines seven through nine, we make the next place prediction for each trace line in the test sequence and compute the prediction accuracy average comparing the corresponding ground truth. We add one to the average if the next place prediction matches with the ground truth and otherwise we add a zero; there is no partial credit for incorrect next place predictions.

### 2.3.2 Periodicity Based Algorithm: PeriodicaB

As we noted before, periodic movements in human mobility are intermixed with outlier cases. We analyze the users’ traces in the MDC data set to verify that the users’ behavior exhibits strong periodic patterns. We compute a user’s visit frequency of places and aggregate total time spent in places and observe regularity in the user’s behavior. As shown in Figure 2.4, which plots the histogram of one typical user’s (user 8 in the MDC data set) visit frequency distribution over places, there are several places dominating the sequence. Visiting time distribution has similar patterns.

Our basic assumption is this: given that the user is currently at place $p_{cur}$, the next place he/she is going to visit should be primarily determined by the location $p_{cur}$ and the time interval in which he/she stays at $p_{cur}$. Therefore, for $p_{cur}$, we check the conditions under which the user goes to a certain next place $p_{next}$. Here, the condition we use is a time distribution $dis.(p_{cur} \rightarrow p_{next})$, which is computed by overlaying all the time intervals in which the user decides to “jump” to $p_{next}$.

Now, if we want to predict the next place after staying at $p_{test}$ at time interval $t_{test}$ (this can be transformed to a time distribution $dist(t_{test})$ the same way as the above), we can use dot product $dist(t_{test}) \cdot dist(p_{cur} \rightarrow p_{next})$ as a similarity measure between the temporal condition at present and the condition of a certain transition. Then we choose the place by maximizing this similarity measure as the next location. Here, we can directly use the dot product without worrying about scaling; this is because the overlay mechanism naturally weighs the distribution by giving the time intervals that correspond to frequent places more weight simply due to the increase in number of visits.

For some places, majority voting (equivalent to a 1st-Order Markov Predictor) will be very efficient. We observed this for places that are not frequently visited. We capture this observation in the algorithm by employing a separate strategy. If the confidence of majority voting is higher than a predefined threshold, we use it; otherwise, we consult the temporal similarity measure.
Observed events with period 20. Events happen at anytime between 20k+5 to 20k+10.

Segment the data using length 20

Overlay the segments

Points are clustered in [5,10] interval

Segment the data using length 10

Overlay the segments

Points are scattered

Fig. 2.3. Intuition of the Periodicity Model from [29]

Fig. 2.4. Visit Frequency Distribution over Places
2.4 Data Sets

We apply *PeriodicB* and *PeriodicaS* algorithms for the MDC and WTD data sets for next place predictions. The MDC and WTD data sets were collected from a diverse group of participants over a long period of time (9 and 4 months, respectively), and the place ID’s are either anonymized or generic access point server numbers.

2.4.1 Nokia MDC Data Set

The data collection campaign for the MDC data set was conducted by Nokia Research Center Lausanne in the Lake Geneva region (Switzerland). Data from smart phones of almost 200 participants was collected in the course of 1+ year. Details of the data set including characteristics of the data set, partition, and availability of different portions for various challenge tasks are described in [30].

Tables in the full MDC data set are shown in Figure 2.5.

2.4.2 Wireless Topology Discovery Data Set

The WTD trace data was collected from approximately 275 freshman PDA users over an 11-week period between September 22, 2002 and December 8, 2002. The freshmen were the initial students residing in the new Sixth College housing facility at the University of California San Diego (UCSD). There was a steady decline in the user population over the trace period due to device-related issues. We have about 90 users with useable trace data to consider for next place prediction by our algorithms.

A background data collection tool was installed on each user’s PDA, and the tool periodically recorded Access Point information as freshmen moved from place to place on the campus. While a student’s device was powered on, WTD sampled and recorded the following information: USER_ID, SAMPLE_TIME, AP_ID, SIG_STRENGTH, AC_POWER, and ASSOCIATED. Overall wireless activity was extensive as students associated with over 400 unique APs in our trace. WTD recorded this information for all APs sensed during a sample; if a user’s device saw three APs in one sample, there were three entries recorded for that sample. For our study, we map an Access Point contacted to by a user as the corresponding location ID for user’s place-to-place transition.

The WTD data set is presented in detail along with analysis of user behavior, activities, mobility and trace-based mobility models in [27]. Users’ locations across the UCSD campus at noon and at 1:00 PM are shown in the Figure 2.6. The dense areas at the lower right are Roosevelt College student housing buildings.

2.5 Mobile Data Challenge by Nokia

Participation in the Mobile Data Challenge by Nokia was the starting point in our journey. We designed our experiments and methods in the context of the challenge submission, and they formed the basis for our next place prediction algorithms. We gained insights into patterns of user movements that are translated into various feature sets for *PeriodicaS* algorithm.
Fig. 2.5. Structure and Tables of MDC Data Set from [30]
Fig. 2.6. Map of the campus with location in WTD Data Set from [27]
Chapter 3

Next Place Predictions

As we discussed in previous chapters, periodic transitions from place to place are inherent in human movements. Through visual examination we detect these periodic movements in traces of user tracking data. However, such user tracking data sets tend to be sparse and incomplete. In addition, periodic movements are surrounded by noise such as transitions to and from less frequently visited places and transitions to one-of-a-kind destinations.

We presented two algorithms in Chapter 2, PeriodicaB and PeriodicaS, that utilize periodicity in movement traces to predict a user’s next location given the current context. Prediction accuracy depends on the ability of the algorithms to recognize patterns of periodic movements while ignoring noise in user traces. The place-to-place transitions used by these algorithms do not contain any semantic information. We deduce and utilize semantic annotations that are inherent in users’ place-to-place transitions in order to improve our prediction accuracy in the next chapter.

We updated PeriodicaB and PeriodicaS from the MDC challenge submission by enhancing the special measures for specific users with unusual periodic behavior. PeriodicaB and PeriodicaS are general purpose next place prediction algorithms that can be used on a variety of data sets gathered from a diverse group of users, with mobility traces collected for a semester or several months.

We use the same data set that was made available for the participants of the Nokia Mobile Data Challenge [30] and the WTD data set from the UCSD Wireless Topology Discovery [27] for running our algorithms. We also present observations from running PeriodicaS on the Reality Mining data set from the MIT Media Laboratory [31].

The aggregated average of next place predictions across all users by PeriodicaS on MDC data set is 49.93%. This average compares well with the average of 52.83% in our submission for the challenge. Aggregated averages from PeriodicaB and PeriodicaS on the WTD data set track the corresponding averages for the MDC data set. We discuss our experiences with the RM data set in Section 3.5.

3.1 Introduction

Reliable predictions are needed for location-based services like recommending nearby establishments to visit and for providing communication services such as data replication and data forwarding. User location is a key determinant for understanding the information requirement needs of the user [32].

Statistical properties of human movements, mobility patterns, and periodic transitions are utilized in generating mobility models for opportunistic networks [33]. We
strive for better understanding of periodicity in user’s captured mobility traces as part
of the process of generating models for our next place predictors.

We illustrate the complexities of detecting and mining periodicity in human mobility through an example: in user tracking traces, we can detect movements with regularity
from an office to several places at noon on weekdays. The person may have lunch in the
office, in a conference room with colleagues, in the office cafeteria, or at an outside establishment. Such lunch time entries will be missing in the data set for a person’s holidays, sick days and business travel days. Semantic annotation of places, user activities and
community affiliations could be used effectively, if such information is either available in
the tracking data set or can be derived from other sources. In these scenarios (given the
current context of office location and timestamp of noon), we can predict with a higher
degree of confidence “lunch place” as the user’s next location.

In the above example, any of four or five locations can be the likely next location
given the current location of office and the timestamp of noon. Even for people with a
high degree of periodicity in their movements, it is hard to reliably predict the user’s
exact next location based on the limited amount of location data in the user’s traces - a next place prediction algorithm, on average, will be able to achieve an accuracy of
only 20% or less in the above example. Our next place prediction algorithms achieve an
aggregated average accuracy of 50%, within a point or two, across all users in the MDC
and WTD data sets.

The rest of this chapter is organized as follows. In Section 3.2 we review existing
work on detecting periodic movements in user data sets and techniques employed in next
place prediction algorithms. We describe the periodic patterns we observed in the data
sets in Section 3.3. Section 3.4 presents the results of the algorithms and our analysis.
In Section 3.5 we present the summary and discuss applications of next place prediction
algorithms in mobility modeling.

3.2 Related Work

Human movements are strongly determined by periodic patterns [34]. Song et.
al. [2] explored the limits of predictability by studying data from anonymized mobile
phone users and found a 93% potential predictability in user mobility. Further, they
discovered that the length of distances covered by users in their mobility patterns has
little impact on predictability.

Location prediction algorithms based on the structure of sequential patterns were
designed using the data mining techniques [35, 36, 37, 38, 39]. These algorithms are
effective when the query for next place prediction consists of the user’s recent mobility
history.

A next place prediction algorithm based on the Mobility Markov Chain that keeps
track of the n previous locations visited was presented in [40]. Our next place prediction
algorithms rely only on the current location and the timestamp, so these techniques
based on the sequential pattern of movements or history of visited places are not useful
for our purposes.

In [41] the dual problems of detecting the periods in movements and mining
periodic movements were studied. Our periodicity-based algorithms are inspired by the
method presented in [29] for detecting periodic behaviors using incomplete observations. We discuss this method in detail through their example on Section 2.2.

3.3 Data Analytics

In this section, we describe the users’ mobility traces, our experiments and insights gained. We use PeriodicaB and PeriodicaS algorithms for making next place predictions. Here we describe the implementation details required to process the mobility traces in the data sets.

The trace records in the MDC data set have a trust flag associated with each place-to-place transition. Our algorithms perform slightly better when considering only “trusted” traces. This is because the untrusted version is more likely to bring noise to the data. The WTD data set does not have a trust flag, and for comparison purposes we did not use the trust flag while using the MDC data set.

We found strong daily and weekly periodicity, as shown in Figure 3.1, for place 3 (the most frequent place) and some weak daily or weekly periodicity for others. This motivates the use of the data folding techniques described in Section 2.2.1 in building user mobility models.

Another interesting observation is if we plot the histogram of places being visited after the most frequent place (place 3) and the histogram of places being visited after one of the least frequently visited places (place 13), as shown in Figure 3.1, we see that after place 3, there are still two places dominating, but after place 13, there is only one place dominating. This observation inspires us to consider a separate strategy for different places. Since we can give only one candidate for the next place when making a prediction, for some places, we may be able to determine the next place simply based on majority voting with high confidence, while for other places we may further explore temporal information to make a decision.

With all the current information of a user, we model our approach into a multi-class classification problem. For each entry, we extract features from the current context and use the next place as the label. We highlight a few issues here. First, the class labels are highly imbalanced since there are dominant places. Meanwhile, many of the minority classes have very few data samples, often only one or two samples. This will cause the classifier to favor the majority classes. Since there are not enough samples for the minority classes, the accuracy on the minority classes will be very low, as we will see later in the analysis.

From the discussions in the previous sections, we see that time and location features are very discriminative in predicting the next place. Thus we only extract features from time and location. While we ignore other context that might be useful, based on our assumption, we avoid possible noise from other information. The features we use here are extracted from: start time of a visit, end time of a visit and current location.

Features. We convert the Unix timestamp in the MDC data set into human readable time according to the time zone. The features that we are interested in contain day of week, hour of day, hour of week, weekend, weekday, morning, noon, afternoon, evening, and midnight. For day of week, hour of day, and hour of week, we use 1-of-K
Fig. 3.1. Frequency distribution of the next place visited after place 3 and place 13
encoding to derive binary values for the features. *Weekend, weekday, morning, noon, afternoon, evening, and midnight*, are themselves binary features corresponding to true or false. We decide whether it is *morning, noon, afternoon, evening, or midnight* by dividing the day time into five time-slots. Each time-slot corresponds to one feature. The bit corresponding to the time-slot that contains the timestamp will be on. Current location is also used as a feature and uses 1-of-*k* encoding.

Up to now, the features we described do not capture any information about the sequence structure. In fact, though, this information sometimes can be very useful. To illustrate, consider a person who moves to a new house during the process of data collection. After the point the user moved, the places the user visits will change, perhaps significantly. Therefore, we add the *normalized start_time* (in the range [0, 1]) as a feature. When testing, we normalize the *start_time* to the range [0.8, 1]. We do not perform this step for the traces in the WTD data set as students did not move to a new house during the data collection period.

**Cross Validation.** All the parameters are tuned via five-fold cross validation. To guarantee fair comparison among different models, the data are partitioned before training, and all the models share the same partition.

**Dealing with unseen places.** We always predict the next place to be a place that the user has visited before. Based on our model, even if the current place is never seen in the training data, we can utilize the similarity in time to make the prediction.

**Experiment Setting.** We conduct all the experiments in Matlab. We implement multi-class classification via support vector machines with the Radius Basis Function (RBF) kernel using the matlab interface of LIBSVM [9].

### 3.4 Predictions and Analysis

Table 3.1 lists the aggregated prediction average accuracies across all users from *PeriodicaB* and *PeriodicaS*. We experimented with different sets of features (with or without the trust flag and splitting into nights and days vs morning/afternoon and evenings) for variations of the *PeriodicaS* algorithm for the Mobile Data Challenge. Using all features, *PeriodicaS* performs slightly better than when we used only the absolute time features. Newly added features (like day of the week, etc.) have some discriminative power intuitively; adding them improves the performance. For users in the MDC data set, classification-based algorithm(s) generally performed better. *PeriodicaB* explicitly emphasizes the temporal similarities in the trace data and, if a person is regular, *PeriodicaB* outperforms the classification-based *PeriodicaS* algorithm.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>MDC</th>
<th>WTD</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>PeriodicaB</em></td>
<td>46.78%</td>
<td>49.04%</td>
</tr>
<tr>
<td><em>PeriodicaS</em></td>
<td>48.93%</td>
<td>45.58%</td>
</tr>
</tbody>
</table>

Table 3.1. Aggregated prediction average accuracies across all users in the MDC and the WTD data sets using the next place predictors *PeriodicaB* and *PeriodicaS*. 
Fig. 3.2. Accuracy of predictions for users in MDC data set
3.4.1 Accuracy Profiles

The aggregated average accuracies listed in Table 3.1 characterize the overall performance of our algorithms and show the amount of periodicity in all user movements. The amount of periodicity in place to place movements varies from user to user, so we analyze the performance of our algorithms on individual users as well.

Figure 3.2 shows the distribution of users in the MDC data set against the percentage of prediction accuracy by the PeriodicaS algorithm. Figure 3.3 shows the distribution of users in the WTD data set against the percentage of prediction accuracy by the PeriodicaB algorithm.

The results show that our algorithm predicts with accuracy in the range of 40% to 70% for the majority of users in the MDC data set. Results for students in the WTD data set are clustered below 50%.

3.4.2 Analysis

The following analysis and figures are generated from the results of PeriodicaS for users in the MDC data set, but the results from PeriodicaB follow a similar pattern. Mobility traces of users in the MDC data sets are captured over a longer period (eleven months) than mobility traces the students in the WTD data set (one semester).

For all users in the MDC data set, we compare the amount of trace data of each user against the next place prediction accuracy for the same user in Figure 3.4. We observe that the individual user prediction accuracy averages vary widely when compared to the aggregated average of 48.93%. We arranged the user IDs on the x-axis in increasing order of their trace file size. We find that average accuracy does not go up for users with higher amounts of trace data. For example, we have twice as much trace data for user #70 in Figure 3.4 compared to user #4 and yet their prediction averages are similar. This confirms that beyond a certain threshold, periodicity of movements does not increase as one collects more user trace data.

We first examine when the models make mistakes. Take user 8 as an example. As shown in Figure 3.5, we plot the accuracy of prediction for each place (i.e. accuracy when this place is to be predicted as the next place) and the accuracy of predictions made from each place. It is clear that the majority classes (frequent places) are usually correctly predicted, while the minority classes are not predicted correctly, and if the user is at a frequent place, prediction of the next place will be less accurate.

We note that there are two cases for the relatively infrequent places. In the first case, the place is only relatively infrequent but still has a good number of samples. In this case, prediction from this place is more accurate. In the second case, the place is very infrequent and only appears once or twice. The prediction from this place will be similar to a random guess.

We also consider user #43 in MDC data set where the predictions have an accuracy around 90%. We find that user #143 has a very compacted histogram of places visited and is highly regular in his/her place transitions, which makes prediction easier.
Fig. 3.3. Accuracy of predictions for users in WTD data set
Fig. 3.4. Accuracy of predictions VS amount of trace data for each user in MDC data set
Fig. 3.5. Error Analysis for User 8
3.5 Discussion

We applied the classification-based algorithm to the Reality Mining data set. The RM data set has trace data from 94 subjects consisting of students and staff at a major university during the months between September 2004 and June 2005. Following are the major items in the data set: a) Phone log with time, description, duration and number, b) Bluetooth log with time and paired devices, c) Location log with time, cell area, cell tower, service provider and user-defined location name.

The Reality Mining data set consists of the observational data as described above and standard self-reported survey data. The observational data is compared with self-reported data, and based on the observational data, friendship network structure was inferred with a high degree of accuracy [31].

We find the aggregated average next place prediction accuracy across all users in the RM data set is 16.11%. Here, the user location is inferred from the Cellular Tower ID. Using the tower IDs and respective transition timings (timestamps when the phone is handed off between cellular towers), phone position was localized to within 100-200 meters. This level of granularity caused ambiguity in movement transitions and severely hampered our periodicity-based models.

There are ways to improve the prediction accuracy for the places that do not have enough personal data. It is straightforward to think about utilizing the correlations between locations using human mobility as proposed by Yu Zheng et al. in [42]. However, one key issue in their method is to find stay point clusters, which they denote as locations that all users share. Then, based on all users’ travel experiences, they detect correlations between locations. In the MDC data set, the locations are user-specific, and we do not perform any location mapping among different users, and nor do we utilize other users’ information to benefit the prediction.

Performance of our algorithms correlates with periodic movements that are present in the data sets. Place semantics, community affiliations, and users’ actions correlating to a given place will improve performance of the algorithms.

Our algorithm predicts with higher accuracy when the test cases do not deal with outlier cases as shown in the analysis of the results. For some applications, next place predictions when a user is at frequently visited places are sufficient. For these cases, mining trace data for frequently visited places and applying the periodicity algorithms achieve higher next place prediction accuracy.
Chapter 4

Semantic Annotations

In the last chapter, we demonstrated the efficiency of PeriodicaS through aggregated average prediction accuracies across all users over two large data sets of diverse participants. PeriodicaS mines periodicity intelligently in users’ mobility traces. In this chapter, we extend the algorithm with additional classification rules. We derive these classification rules by applying explicit semantic annotations (home, work place and public transportation points associated with places visited), and accompanying group information.

We propose novel ways of transforming bits of information in the mobility traces, defined to be inherent semantic annotations, as features for mobility modeling. Inherent semantic annotations are computed with temporal variations from visited places such as end time only and duration time only. We deduce more inherent semantic annotations from place rankings by frequency of visits. By progressively employing these two types of semantic annotations, explicitly stated in the data set and deduced from the mobility traces, we improve next place prediction accuracies up to 54% compared to baseline predictions.

4.1 Introduction

Next places are commonly predicted using standard data mining techniques based on collected mobility traces. Given that these traces are collected over a long period of time (approximately a year), they contain noise and tend to be incomplete. As many factors are involved in mobility [43], predicting a user’s next place accurately and consistently is a vexing problem.

When predicting next places with only the corresponding GPS coordinates and no benefit of semantic information, prediction accuracies, as illustrated by [44], are in the neighborhood of 50%. This was demonstrated as the best result in the global challenge of next place prediction algorithms from the submissions of 79 participants. By using periodicity in user movements and utilizing available semantic information, prediction algorithms can achieve higher accuracies. For example, based on a user’s mobility traces, we can predicting lunch place as the user’s next place more accurately than a specific restaurant. Typically, the user will visit many different restaurants during the time period of collected traces.

In this chapter we present results that show how different types of semantic annotations of a location impact the ability to accurately predict human mobility. We start with PeriodicaS, which utilizes periodicity to predict next locations. We also consider explicitly stated and deduced inherent semantic annotations to quantify their impact on making accurate predictions.
We bring together periodicity detection techniques and the practice of mining semantic annotations as a powerful combination for improving users’ next place predictions. We study users’ traces to glean their mobility patterns through the semantic annotations of the visited places. The semantics may be explicitly stated in the data set or deduced. We examine specific time segments, durations and frequency of place visits. We then translate these attributes into features for training and making predictions in our algorithm.

Mobility traces capture a user’s place-to-place transitions over a period of time; each transition is represented as a line item in tables of a data set. These transitions, as shown in Figure 2.1 are converted to a feature vector (current place, start time, end time) and a given a corresponding classification label (next place). PeriodicaS builds a mobility model independently for each user and makes next place predictions based on this model.

Next place prediction algorithms typically improve their accuracies by taking advantage of community information such as membership and meeting times and places. Mobility traces in our data set are user-specific, and there are no defined communities among the participants. We focus on place semantic and demographic information to understand the users’ mobility patterns and enhance individual users’ next place predictions.

We describe the details of PeriodicaS, the data set and our methodologies for training and testing data sets in Section 4.3. Section 4.4 describes prediction improvements obtained by PeriodicaS using the collected place semantic annotations in the data set and user demographic information; we refer to these as explicit semantic annotations. We deduce inherent semantic annotations through: a) omitting start time, end time and other temporal variants in our feature vectors, b) segmenting traces by weekdays and weekends, and c) using visited place frequencies, duration times and other spatial considerations. We call these deductions inherent semantic annotations and describe the corresponding impact in prediction accuracies in Section 4.5. We conclude the chapter with a summary and discuss a technique for combining several place-to-place transitions into a denser region for improving next place predictions in Section 4.6.

4.2 Related Work

Mobility traces in an indoor setting are considered in [21]. They define a test bed for collecting data and a framework for analysis. In the controlled environment of a campus, [20] predicts both current place residency times and next place predictions by combining time-of-day features with an additional feature of users’ group size as users transition from place to place. Ye et al. present semantic annotation algorithms for places in location-based social networks in [45].

Techniques for mining attributes effectively in large, incomplete and sparse data sets with noisy traces were illustrated in [46] and [41]. We follow the method proposed in [29] for clustering user transitions into periods based on daily and weekly patterns.
4.3 *PeriodicaS* and Feature Sets

We started exploring user mobility traces with participation in the Nokia Mobile Data Challenge (https://research.nokia.com/page/12000/). Details of the data set, including characteristics of the data set and various challenge tasks, are described in [30]. We used the MDC competition data set for the experiments in Chapter 3. We use the full data set, available from the Mobile Data Challenge Initiative (https://www.idiap.ch/project/mdc/), for all the experiments discussed in this chapter. The full data set contains traces for 152 users, whereas the data set made available for the competition consisted of 90 users.

*PeriodicaS* follows the commonly used steps in the Knowledge Discovery in Databases (KDD) process. We use a pre-processing step to convert each place-to-place transition to a set of attributes: start day of week, end day of week, start hour of day, end hour of day, start minute, end minute, current place ID, normalized start time, and the users’ next place. Underpinnings of the *PeriodicaS* algorithm are detailed in [43].

In the data mining step, we convert transition attributes into a feature vector and establish the next place as the corresponding label for the feature vector. We vary the usage of these attributes in different ways in building user mobility models, as described in the following sections.

We use LIBSVM (Library for Support Vector Machines (SVMs)), as described in [9] for the classification task and model generation.

As part of the results validation step, we use the `svmpredict` function and compute the accuracy of next place predictions as described in the previous section.

*PeriodicaS* is implemented in MATLAB and uses the MATLAB/OCTAVE interfaces of LIBSVM.

<table>
<thead>
<tr>
<th>Place Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Home</td>
</tr>
<tr>
<td>2</td>
<td>Home of a friend, relative or colleague</td>
</tr>
<tr>
<td>3</td>
<td>Work place or school</td>
</tr>
<tr>
<td>4</td>
<td>Location related to transportation</td>
</tr>
<tr>
<td>5</td>
<td>Workplace or school of a friend, relative or colleague</td>
</tr>
<tr>
<td>6</td>
<td>Place for outdoor sports</td>
</tr>
<tr>
<td>7</td>
<td>Place for indoor sports</td>
</tr>
<tr>
<td>8</td>
<td>Restaurant or bar</td>
</tr>
<tr>
<td>9</td>
<td>Shop or shopping center</td>
</tr>
<tr>
<td>10</td>
<td>Holiday resort or vacation spot</td>
</tr>
</tbody>
</table>

Table 4.1. Place Labels and Descriptions

From our results as described in Chapter 3, we established that *PeriodicaS* predicts the user’s next place well. Here, we explore adding different semantic annotations,
Fig. 4.1. Baseline Prediction Accuracies for all Users

Mean = 50.47  
Median = 47.53
as [47] shows that simple classification rules improve prediction accuracies in commonly used data sets.

4.3.1 Preprocessing Mobility Traces

As the user trace data spans more than nine months, we assign a weight (normalized start time) to give more importance to the transitions that occur later in the data collection campaign. We divide the training data sets into five partitions (each transition is randomly assigned a partition number) for cross validation. We discuss the logistics of dividing the user traces for training and testing purposes and provide a description of aggregated accuracy averages.

After predicting the next place for each of the traces in the test sequence, we compute the prediction average accuracy for the user. Predictions matching the actual next place in the test sequence scored as one and the predictions not matching scored as zero in computing the average. The aggregated prediction accuracy (referred to as the final score in judging MDC submissions) is the average prediction accuracy across all predictions for all the users in the data set.

Some users may go through life-changing experiences with respect to living arrangements, changes in work place locations, and seasonal routines. These changes, reflected in the traces of a user, will negatively impact the prediction accuracies.

We divide the traces in two different ways: a) 80Split20 - the first 80% of the traces for training and the most recent 20% of the traces for testing and b) 80Random20 - randomly remove 20% of the traces for testing leaving the remaining 80% of the traces for training. We consistently observe that models trained and tested using the 80Random20 method produced accuracies in the range of 2% to 5% better as compared to the models trained and tested using the 80Split20 method of dividing the traces. This is in line with our expectations, as the risk of a major change impacting the traces is more evenly distributed with the 80Random20 method of splitting the traces. In the rest of the chapter, we present the results using the 80Random20 method only.

We determine training accuracies and testing accuracies in the PeriodicaS algorithm during the corresponding phases of training and predicting. Testing accuracies correlate with the training accuracies in all the cases. However, the training accuracies are consistently higher by 5% to 15%. We always present only the testing accuracies, not the higher training averages.

We ran PeriodicaS on the full data set to compute aggregated prediction accuracies across all users, and the results are shown in Figure 4.1. These results compare well with our submission results, validating PeriodicaS with the full data set. For the rest of the chapter we refer to these results as the baseline.

4.4 Collected Semantic Information

In this portion of our study, we take into consideration the semantic information (labels) associated with places visited by the user. Place label information collected from user mobility traces and captured in the data set is described in Table 4.1. Place labels describe the kind of place: user’s home, home of a friend, relative or colleague and so
on. All places in trace files that do not have the label information were mapped into a single composite place designating that there is no semantic information available for that place. Place label information is also user-specific; for example, Place Label 4 in Table 4.1, a location related to transportation, may not be shared among users.

4.4.1 Accuracies with Place Labels and Companion Information

Associated with place labels is companion information for visited places in the traces. The participants’ transitions from place-to-place have been tagged with family, with close friends, with friends, and with colleagues or acquaintances. The labels are captured along with the traces, the places are tagged with group accompanying information in the data set, and the places that do not have any group accompanying information are tagged as incidental.

Many users have only three places with label information (home, home of a friend and work place) in their traces. Thirty-six participants do not have any place label information; these users are not considered when running this experiment. We show the aggregated average prediction accuracy with label information and with label plus companion information in Figure 4.2. The average prediction accuracy improves by 48% with place labels as compared to the baseline accuracy. We see a similar improvement when companion information is combined with place labels, as shown in Figure 4.2.

We observe that the group accompanying tags add minor information to the semantic annotations of place labels and improve the prediction accuracies by less than 1%.

4.4.2 Demographics

The data set contains demographic information from the participants. We consider these information bits as explicit semantic annotations. We do not apply this information as a set of features for our mobility modeling. Instead, we segregate the users by their demographic information, run PeriodicaS on segregated groups and report the results. While the mobility traces are recorded by the smart phones automatically and are uploaded to the data server, the demographic information was supplied by the participants through filling out forms.

The demographic information (gender, age group, number of people in the household, work status and role of the user paying the phone bill) is extracted from forms filled out by the participant.

For the next set of experiments, we select subgroups of users based on the demographic data described above, and we use the trace files with place label information from the previous section. Subgroups of male students age 16 and up (15 participants), female students age 16 and up (23), male workers age 22 and up (20), and female workers age 22 and up (50) are used for calculating the aggregated average accuracies.

We present the aggregated average accuracies of next place predictions from PeriodicaS for these four subgroups in Figure 4.3. Aggregated prediction accuracies improve by 22% to 40% for different groups of participants.

We ran PeriodicaS with more subgroups based on work status and age. The average prediction accuracies compare to the averages presented in Figure 4.3. For
Fig. 4.2. Aggregated average accuracies (as compared to the baseline of 50.47%) with place labels and group accompanying tags.
smaller, more focused subgroups, we have observed higher prediction averages. For example, the average prediction accuracy for the group of male workers between the ages of twenty eight and thirty three (total of twelve participants) is 79%. We notice that the aggregated averages for student groups are 10% less than corresponding working groups, as students have less periodicity in their movements.

The aggregated averages for some of the demographic groups will be higher when the inherent semantic annotations from the following section are applied.

4.5 Inherent Semantic Annotations

One of our main contributions is recognizing the inherent semantic annotations in basic place-to-place transitions and harvesting them to improve next place predictions. This allows us to improve next place predictions for a user through a straightforward periodicity mining algorithm like PeriodicaS.

The inherent semantic annotations are deduced by a) manipulating temporal variants in the traces, and b) accounting for spatial considerations of place ranking and incidental place.

4.5.1 Temporal Variants

Mobile traces contain start time and end time associated with a place before a user’s transition to the next place. In this section we experiment by varying these timing constraints as part of the feature set for modeling user mobility. We use these feature sets for training and predicting the user’s next place and present the prediction accuracies.

In the baseline test case we take into account both the start time and the end time in place-to-place transitions. We consider feature sets with end time only and with no timing features, i.e., only place-to-place transitions.

We compute durations of stay for each place by subtracting the start time from the end time in the user’s transitions. We then replace the start/end times by duration of the place for both training and testing (predicting).

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Aggregated Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 End time (no start time)</td>
<td>51.47%</td>
</tr>
<tr>
<td>2 Place to place (no time constraints)</td>
<td>47.63%</td>
</tr>
<tr>
<td>3 Duration at a place (no start/end time)</td>
<td>49.14%</td>
</tr>
<tr>
<td>4 Baseline</td>
<td>50.47%</td>
</tr>
</tbody>
</table>

Table 4.2. Prediction Accuracies with Temporal Variants

Prediction accuracies for three experiments along with the baseline are listed in Table 4.2. We observe that the average, when considering the end time of transitions, is comparable to the baseline. As we have a diverse set of participants with different
routines for work and other day-to-day activities, the impact varies in next place predictions for different users. This issue is highlighted in the experiment of place-to-place transitions with no timing considerations. There may be frequent transitions from place-$i$ to place-$j$ and place-$k$, with no time information available. In this case the prediction algorithm defaults to majority voting, which is not the best way to take advantage of users’ periodicity in their place-to-place transitions.

4.5.2 Place Ranking

The number of unique places visited by users during the data collection campaign varies widely, from 15 to 195 places. Figure 4.4 shows the distribution of unique places visited by all the users in the data set. On average participants in the data set visited 71 places during the data collection campaign. We are interested in studying the users’ mobility patterns by exploring frequency of visits and time spent at each of these places.

First we describe the distribution of the top ten frequently visited places for all users and the notion of an incidental place. By scanning the mobility traces, we rank places by frequency of user visits. The frequency of visits drops off drastically for all places after the two or three most frequently visited places as illustrated in Figure 4.5. We set a threshold of number of most frequently visited places, and we combine all the infrequently visited places in a user’s trace into one composite place called the incidental place. Determination of all incidental places is user specific, as is the case with all the places in our mobility traces. We keep the complete trace history intact except for replacing all the infrequently visited places with the incidental place.

Figure 4.5 shows the place distribution for all the users with the top 10 most frequently visited places plus the incidental place. The visit frequency of the most dominant place for all the users is higher than 40%. The frequency with which a user visits an incidental place depends on the combined frequency of the top two or three dominant places and the number of unique places the user has visited during the campaign. We observe similar distributions when we rank places by the time the user spends at the visited places.

We rank places by duration, time spent at a place, and visit frequency. We use these rankings as the inherent semantics in our attributes-to-features translation.

Mapping of the incidental place produces the following advantages, in addition to reducing the number of places for next place prediction:

- Typically the infrequently visited places are not of interest, as we may not be able to provide services to the user at these places.

- We observe that users often transition from one of the dominant places to the incidental place and immediately back to one of the dominant places. With the incidental place in user traces, the probability is higher for accurate predictions to the dominant place.

- Generating mobility models with two sets of traces – with and without incidental place mapping – allows us to make a second prediction of the actual infrequently visited place after the incidental place is predicted.
We ran experiments with the place rankings of the top 10 and the top 15. The aggregated averages are close to the baseline average. We present the aggregated averages considering the users’ traces of top 3, top 5, and corresponding incidental places in Table 4.3 with two different classification rules described in the table.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Description</th>
<th>Aggregated Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration 3</td>
<td>Top 3 places by time spent</td>
<td>59.15%</td>
</tr>
<tr>
<td>Duration 5</td>
<td>Top 5 places by time spent</td>
<td>55.09%</td>
</tr>
<tr>
<td>Frequency 3</td>
<td>Top 3 places visited</td>
<td>58.32%</td>
</tr>
<tr>
<td>Frequency 5</td>
<td>Top 5 places visited</td>
<td>54.60%</td>
</tr>
<tr>
<td>Baseline</td>
<td>Traces with all the places</td>
<td>50.47%</td>
</tr>
</tbody>
</table>

Table 4.3. Place Ranking and Prediction Averages

There are four candidates (top 3 places plus the incidental place) for predicting the next place in the cases of Duration 3 and Frequency 3. In these cases, aggregate prediction averages improve by 20% compared to the baseline average as shown in Table 4.3. We observe that the frequency of visits to the top 3 places of many users correlates with visit frequencies of place labels (home, friends’ home and work place) as described in Section 4.

4.6 Summary and Discussion

We validate that PeriodicaS works well as a next place prediction algorithm for mobile traces of the general population. The accuracy of predictions correlates well with periodicity in the user’ movements. Considering only the place-to-place transitions, PeriodicaS achieves a baseline aggregated average accuracy of 50.47%.

The averages improve by 48% by adding semantic information of place label information. In addition to utilizing the label information, we aggregate prediction averages for smaller groups based on users’ demographic information: all working male, all working female, all male students and all female students in certain age groups. The aggregated averages further improve in the range of 15% to 25% for the smaller groups.

We devise a novel mechanism for deducing inherent semantic annotations from the users’ basic place-to-place transitions by considering the temporal variations in their traces and ranking their visited places. Taking advantage of these inherent semantic annotations, PeriodicaS improves the averages by 20%.

The MDC full data set contains tables of system information, social communication, user agenda/contacts, phone usage, networking (gsm, bluetooth, wlan), location and movement, shared values, and meta data on user generated information. Semantic annotations can be deduced from this information and used for improving prediction averages.

We observe that some users have dense regions in their mobility traces; these dense regions can be explored and, where appropriate, replaced by a single place. We can link major paths and the most important routes among these regions. This gives us
the ability to view mobility traces in a hierarchical network model by examining the trace data at a different level of granularity. In this scheme the incidental places and their associated traces will be absorbed into other (significant) place-to-place transitions using the technique of link predictions \cite{l2}. Thus we can utilize the users’ mobility network with a smaller number of places by combining several transitions. We can also take advantage of the semantic information, when available, in constructing such mobility networks, and the aggregate prediction averages will improve accordingly.

General data sets containing mobility traces from diverse participants exhibit less periodicity compared to trace files of a focused user group, such as those working on a mission critical campaign/project, students belonging to an honors college participating in an accredited academic program, or a group of athletes on a sports team. Participants in such a study will have better periodic movements, and in such cases next place prediction accuracy will be higher, as we demonstrated with the users’ demographic information in our data set.
Fig. 4.3. Prediction accuracies for four different groups; baseline average is 50.47%
Fig. 4.4. Number of unique place IDs visited by users
Fig. 4.5. Distribution of top 10 places visited
Chapter 5

Mobility Forecasting

PbMFS - Periodicity based Mobility Forecasting System - is built around our next place prediction algorithm PeriodicaS. This paper describes the context, components, structure, operational flow and benefits of PbMFS. We illustrate the benefits of PbMFS through generating end-to-end trajectories of user mobility and forecasting users’ movements involving to and from places of interest. There are two ways to generate mobile trajectories for an individual user with the help of a next place predictor: a) based on individual user mobility models built from their mobility traces - sequence of place-to-place transitions with time stamps - and b) based on the aggregated time spent at places of interest by the user. Making accurate next place predictions through the former approach is beneficial but more difficult. PeriodicaS follows the first approach, and the performance of the algorithm is characterized in [43]. Users’ mobility traces contain a large number of places, and the next place prediction algorithms may not have sufficient amount of reliable trace history to select a next place destination from least frequently visited places. We demonstrated in [49] that accuracy of next place predictions increases as we consolidate the number of places visited by the user in mobility trace files through the inherent semantics of the visited places. Deduction and effective usage of inherent semantics is the distinguishing characteristic of our forecasting system. While forecasting a user’s mobility trajectory, PbMFS dynamically chooses the appropriate feature set for a given next place prediction and switches between feature sets during the sequence of predictions as required.

5.1 Introduction

Service providers need to anticipate users’ needs, desires and actions to improve their services to the user, and understanding user mobility and anticipating their next place prediction(s) is an important step in that process. For example, the ability to anticipate user mobility can improve functionality of digital assistants, such as Siri and Cortana, significantly [50] [25].

Accurately predicting a person’s next place destination is a hard problem, especially in the absence of visited places’ semantic information. Several techniques and algorithms as point solutions exist in the literature [44]. Predicting a user’s next place destination based on place-to-place transitions is much harder than making next place predictions, considering the distribution of time spent by users at frequently visited places. These two variations of predicting a user’s next place – using current time only versus using the current time and current place – are explored in [22].
Our next place predictions are always based on place-to-place transitions with current place and time. We use the sequence of predictions for generating a user’s end-to-end mobility trajectory for a given time period.

We describe the components and operational flow of $P_bMFS$, Periodicity based Mobility Forecasting System. $P_bMFS$ is a versatile system that forecasts users’ future place-to-place movements based on their available mobility traces. We co-opted PeriodicaS, our proven next place predictor, as the core component of $P_bMFS$. Periodicity and place semantic annotations in users’ mobility traces are leveraged by PeriodicaS for accurately predicting next place destinations. We deduce inherent semantic annotations and convert these annotations into a rich set of feature vectors.

We illustrate its power through the process of generating end-to-end trajectories of user mobility and forecasts of users’ movements involving “to” and “from” places of interest. In [49] accuracy of next place predictions increases as we consolidate the number of places visited by the user in a trace file through the inherent semantics of the visited places. Deduction and effective usage of inherent semantics is the distinguishing characteristic of our forecasting system.

We start with mobility traces collected over a period of time, recognize existing patterns, exploit properties of periodicity, and deduce inherent semantic annotations to predict next place destinations. Using these predictions we build users’ end-to-end trajectories, which form the base models for service-providing opportunities.

5.2 Related Work and Background

There is a large amount of existing work in modeling user mobility. This class of work is based on mathematical models of mobility. For example, mobility models are generated and characterized using well known techniques such as Random Walk [12], Random Waypoint [13], Reference Point Group Mobility (RPGM) [14] [15], and Manhattan [16].

A second class of mobility models are generated based on known specific patterns. For example, the Universal Mobility Modeling Framework (UMMF) [17] generates specific mobility models based on user-supplied parameters. With this framework, and based on Repeated Traversal, Bounding Overwatch, and Pincer Movement military strategies, tactical mobility models can be generated [18]. Using the models generated from these studies, data replication and other communication service algorithms are analyzed in [19].

Neither of these types of models are useful for the practical purposes of providing communication services. In comparison, we take the approach of mining patterns from collected mobility traces and build models based on forecasted user mobility trajectories.

Periodicity is defined as a quality, state, or fact recurring at regular intervals, such as every 24 hours. We are interested in periodic patterns that exist in human movements that are part of individuals’ daily mobility history. In [2], Song et al. found that there is 93% potential predictability in user mobility. They studied two data sets, an anonymized data set representing 14 weeks of call patterns from 10 million mobile phone users, and a second set of anonymized records of 1000 users, whose coordinates were recorded every hour over eight days. Despite the inherent population heterogeneity
in their study, they found the maximal predictability varies very little between a high of 93% and a low of 80%. Their work provides the framework for characterizing periodicity in a user’s mobility patterns while we demonstrate suitable methods for effectively using periodicity for next place predictions and mobility forecasting.

In this chapter we use a mobility data set that was collected by the Nokia Research Center Lausanne in the Lake Geneva region (Switzerland) as part of the Mobility Data Challenge (MDC). Data from the smart phones of almost 200 participants was collected in the course of 1+ year. Details of the data set including characteristics of the data set, partition, and availability of different portions for various challenge tasks are described in [30].

After training with 80% of the user mobility traces, we use the rest of the 20% of the traces as the ground truth for checking the accuracy of our next place predictions. We use LIBSVM [9] for modeling, training, and making next place predictions.

- Given a user’s current context of place ID and timestamp, `svmpredict` from LIBSVM determines the decision values or probability estimates to determine the next place prediction. These probability estimates are internal to LIBSVM; we get the next place prediction place ID as a returned value.

- We reserve 20% of the traces from modeling and training by LIBSVM. We make next place predictions for each of these traces and compare the results. We assign a one when the predicted next place matches with the ground truth and otherwise we assign a zero for computing the average accuracy for a sequence of predictions. The average of these predictions gives us the accuracy rate or confidence level in making next place predictions for a given user.

- We aggregated next place prediction accuracy averages across all users. This aggregated average is a good indicator for the performance of PeriodicaS at a group level in making next place predictions using a given feature vector.

PeriodicaS is a general purpose next place predictor developed using standard KDD processes [13]. PeriodicaS does not depend on group mobility properties or community activities, as that information is not readily available in the trace data sets. We exclusively deal with aggregated averages and compare them with other aggregated averages to measure the benefit of different features sets. We require a better understanding of individual user prediction accuracies to characterize a single user’s mobility forecasts. These user-specific mobility forecasts are candidates for developing models for user’s specific opportunistic networks. The participants in our data set do not form a cohesive group and they do not belong to any type of social group nor community.

Section 5.3 starts with formal representation of user mobility traces and an illustration of temporal and spatial characteristics embedded in basic place-to-place transitions, which are collected over a period of time and are provided as users’ mobility traces in data sets. We expand the concept of inherent semantic annotations, introduced in [49], and present the aggregated accuracy averages of next place predictions for all users with our next place predictor PeriodicaS with all available place annotations.
We deduce inherent semantic annotations based on distance matrices in Section 5.4. \textit{PeriodicaS} generates the best next place predictions using the clusters of visited places computed based on these inherent semantic annotations.

We systematically generate multiple choices for any next place prediction in Section 5.5. With primary and backup predictions \textit{PeriodicaS} meets the high water mark of 93% potential predictability of user mobility from \cite{2}.

In Section 5.6, we analyze next place predictions of individual users that constitute the aggregated averages presented in the other sections. We present our mobility forecasting system, $P_{bMFS}$, in Section 5.7, show its operational flow and and generate different types of user mobility forecasts.

5.3 Place-to-place Transitions and Annotations

Best of the class next place prediction algorithms, based on place-to-place transitions without semantic information, achieve an aggregated average accuracy of 56\% \cite{44}. Next place predictors need more information associated with visited places to improve prediction accuracy.

Additional information may be collected through forms and indirect methods. While this information may not be available in real-time, it can be used to improve the forecasting system.

5.3.1 Visualization

Human mobility models are studied in \cite{11} through the generation of graphical models. Mobility network graphs bring out the structural information from the users’ place-to-place transitions. These visual representations illustrate the difficulty in recognizing and exploiting a user’s mobility patterns in making next place predictions.

We show the graphs for a user (#39) from the MDC data set in Figures 5.1 and 5.2. User #39 visited 128 different places during the 12-month period the trace data was collected. We collapsed the time stamp information of all the place-to-place transitions in the graph of Figure 5.1. These transitions are represented as a single edge in Figure 5.1. The thickness of the edge from a given node is in proportion to the number of visits to and from the corresponding place ID. The raw mobility graph with one edge per each place-to-place transition in the trace file for this user consists of 12,374 edges.

The graph in Figure 5.2 shows all the place-to-place transitions among the top ten frequently visited places. We combined the infrequently visited places into a composite place ID: \textit{incidental place}, 11 in Figure 5.2. Place-to-place transitions from place $i$ (a top 10 frequently visited place) to place $j$ (not in the top 10 frequently visited places) are now part of the edge from the node representing $i$ to the node representing \textit{incidental place}.

5.3.2 Existing Annotations

The Nokia MDC full data set \cite{30} contains tables of User Forms along with System Information, Phone Usage, Social Communication, and Networking. Parts of the
Fig. 5.1. Mobility Traces of user #39 with all 128 places
information in these tables *directly* annotate the place-to-place transitions semantically and are helpful for our next place predictor.

Tables of User Forms consist of the label information of the places visited by the participants. Label information was computed after the data acquisition campaign based on the *gps* information. Demographic information in these tables was extracted from the forms filled out by the participants.

We employed our previously published algorithm, *PeriodicaS* [13], using place labels, companion information and the demographics of users. Techniques for extracting this information and methods to map this information to appropriate feature vectors were introduced in [49]. *PbMFS* takes full advantage of this information with each invocation of *PeriodicaS* during the process of forecasting a users’ mobility trajectory.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Aggregated Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Place Labels</td>
<td>69.10%</td>
</tr>
<tr>
<td>2 Companion Information</td>
<td>70.41%</td>
</tr>
<tr>
<td>3 Female Workers age 22 and up</td>
<td>77%</td>
</tr>
<tr>
<td>4 Male Students age 16 and up</td>
<td>69%</td>
</tr>
<tr>
<td>5 Baseline (no semantic annotations)</td>
<td>50.47%</td>
</tr>
</tbody>
</table>

Table 5.1. Prediction Accuracies Utilizing Existing Annotations from the Data Sets

We summarize the aggregated average next place prediction accuracies from *PeriodicaS* with different annotations in Table 5.1. Next place prediction averages improve up to 54% compared to the *baseline* average.

5.3.3 Inherent Semantic Annotations

Aggregated average prediction accuracies across all users are good indicators for the overall performance of next place predictors. We experiment with various inherent semantics and determine their usefulness through comparing corresponding aggregated accuracy averages.

The concept of inherent semantics was introduced in [49]. We extend the concept in this section. Table 5.2 lists the description of a feature set, corresponding abbreviation, and the group average. The baseline aggregated average across all users in the dataset is 50.47%.

The aggregated average accuracies of next place predictions of weekdays (WekS) and all seven days (*Baseline*) are similar - 51.05% vs 50.47%. Prediction accuracies for weekends (WeeS) are lower by 10% compared to the baseline average, as there is less periodicity in user movements during weekends.

When we train and predict the next place transitions for a particular weekday (Wednesday), we lose the benefit of a user’s history of movements from all other weekdays in training our models. We explored the case of training with transitions from all weekdays with different weights while predicting next places for Wednesday as a way of...
Fig. 5.2. Mobility Traces of User #39 with the Top Ten Frequently Visited Places and the Rest of the Places Combined into an Incidental-place (11)

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Abbreviation</th>
<th>Aggregated Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weekends</td>
<td>WeeS</td>
<td>44.64%</td>
</tr>
<tr>
<td>Place to Place</td>
<td>PtPS</td>
<td>48.19%</td>
</tr>
<tr>
<td>Duration</td>
<td>Durs</td>
<td>49.14%</td>
</tr>
<tr>
<td>Weekday (Wednesday)</td>
<td>WedS</td>
<td>49.44%</td>
</tr>
<tr>
<td>Baseline (no annotations)</td>
<td>BLS</td>
<td>50.47%</td>
</tr>
<tr>
<td>Weekdays</td>
<td>WekS</td>
<td>51.05%</td>
</tr>
<tr>
<td>End Time</td>
<td>EndS</td>
<td>51.47%</td>
</tr>
<tr>
<td>Top 5 by Frequency</td>
<td>T5FS</td>
<td>54.60%</td>
</tr>
<tr>
<td>Top 5 by Duration</td>
<td>T5DS</td>
<td>55.71%</td>
</tr>
<tr>
<td>Top 3 by Frequency</td>
<td>T3FS</td>
<td>58.07%</td>
</tr>
<tr>
<td>Top 3 by Duration</td>
<td>T3DS</td>
<td>59.15%</td>
</tr>
</tbody>
</table>

Table 5.2. Prediction Accuracies Utilizing Inherent Semantic from the Trace Data
preserving the history and obtained mixed results. This approach of weighting place-to-place transitions discriminatingly instead of making inclusion or exclusion decisions with respect to transitions in the training phase is promising for improved prediction accuracies and deserves further study.

We deduce inherent semantics based on the rankings of the place IDs by visit frequencies and time spent. As a background process of $P_bMFS$, we ran PeriodicaS with different inherent semantics and compile a database of feature vectors and corresponding aggregated prediction accuracy averages. While generating mobility forecasts, $P_bMFS$ determines the right feature set to use based on the context of the query and the accuracy average for a given user.

### 5.3.4 Most Frequently Visited Places

In this section, we study the mobility patterns of users involving their most dominant places by visit frequency. We present two cases of PeriodicaS making next place predictions with a) two most dominant places and a composite incidental place, and b) the most dominant place and a composite incidental place representing all the other places. Though the average number of places visited by a participant during the data collection period is 65, we observe that for many users there are two or three dominant places, by visit frequency, in their mobility traces. Figure 5.1 shows that User #39 visits three places most frequently out of the total of 128 places.

Understanding the performance of the next place predictor is important, as we often make user’s next place predictions “to” or “from” their most frequently visited places. We use these cases effectively in generating user mobility forecasts involving any two visited places of interest.

The average accuracies of next place predictions for all users in the case of considering the two most dominant places visited are shown in Figure 5.3. The aggregated next place prediction accuracy average of 71.35% conforms to the pattern of increasing aggregated accuracy averages (Baseline average of 50.47% and the average with three most dominant places is 59.15%).

Figure 5.4 presents the prediction average accuracies for all users when considering only the most dominant place visited. As the incidental place now represents all the remaining places, there is a higher probability for transitions from the incidental place back to itself. There are two equally probable choices for next place destination from the incidental place a) to the most frequently visited place and b) to the incidental place itself. A random next place predictor will have an aggregated average accuracy of 50%. With PeriodicaS, aggregated prediction averages across all users improve by 60% compared to the baseline.

While running PeriodicaS with arbitrary place IDs instead of the most dominant visited place IDs of users, we observed the aggregated next place prediction average accuracies to be similar. We predict the incidental place as the destination in case of an arbitrary place ID ($j$) instead of the dominant place ID. This becomes a negative prediction with respect to place $j$, and such next place predictions with high accuracy are useful for (re)deploying communication services.
Fig. 5.3. Prediction Accuracies with Two Most Dominant Places - Average Accuracy is 71.35% as Compared to the Baseline Average of 50.47%

Fig. 5.4. Prediction Accuracies for the Most Dominant Place - Average Accuracy is 81.77% as Compared to the Baseline Average of 50.47%
5.4 Inherent Semantic Annotations based on Distance Matrices

As noted earlier, there are patterns in users’ place-to-place transitions. We observed that proximity of places is a dominant factor in users’ mobility traces.

Reducing the number of visited places improves accuracy. We showed the methods to categorize the visited places effectively for improving the efficiency of our next place predictor, PeriodicaS, through inherent semantic annotations. We derive spatial inherent semantic annotations based on distance matrices of visited places and construct logical clusters of the visited places.

Visited places are user-specific, and the place IDs are assigned sequentially in the order of first visit to the places. The visit sequence for each user starts with place ID 1.

The absolute coordinates of places visited in the place-to-place transitions are hidden. A coarse distribution matrix between visited places was computed for each participant during the data collection campaign. Computed distance matrices provide more reliable and accurate information about the user-visited places than the information collected through the user-completed forms and other methods.

The matrix of relative spatial distances between places is descretized into four levels as shown in Table 5.3. We extract these distance matrices from the MDC challenge data set, build clusters of visited places, and make next place predictions based on resulting clusters of places.

<table>
<thead>
<tr>
<th>Level</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Less than 1km</td>
</tr>
<tr>
<td>2</td>
<td>1-5km</td>
</tr>
<tr>
<td>3</td>
<td>5-10km</td>
</tr>
<tr>
<td>4</td>
<td>10km+</td>
</tr>
</tbody>
</table>

Table 5.3. Relative Spatial Distances between Places

In this section, we describe the Greedy and Hierarchical clustering algorithms, derive clusters, and present the aggregated next place prediction accuracy averages. PeriodicaS achieves the best next place predictions using the clusters formed through the hierarchical clustering algorithm.

5.4.1 Clustering Visited Places

Our first attempt at clustering places is based on the greedy approach using the proximity of visited places. This approach is described in Algorithm 2.

The greedy algorithm takes the distance matrices as described in Table 5.3, number of visited places, and desired number of clusters as inputs. Our greedy algorithm constructs the clusters based on an optional seed vector, and clusters are formed around specific visited places. By default, the greedy algorithm starts with no seeds and forms clusters centering around visited places with the most nearest neighbors.
Algorithm 2: Greedy Algorithm

Input: number of place IDs \( n \), distance matrix \( M \), seed vector \( X \), and number of clusters \( k \)

Output: clusters of place IDs, \( P : \{P_1, P_2, \ldots, P_k + 1\} \)

1. for \( c \leftarrow 1 \) to \( k \) do
2. \[ \text{max} \leftarrow 0; \text{index} \leftarrow -1; \text{count} \leftarrow \{0, 0, 0, 0\} \]
3. \[ \text{for } i \leftarrow 1 \text{ to } n \text{ do} \]
4. \[ \text{if } X[i] \neq 0 \text{ then} \]
5. \[ \text{for } j \leftarrow 1 \text{ to } n \text{ do} \]
6. \[ \text{if } M[i, j] = 1 \text{ and } X[j] \neq 0 \text{ then} \]
7. \[ \text{count}[i]++; \]
8. \[ \text{if } \text{count}[i] > \text{max} \text{ then} \]
9. \[ \text{max} \leftarrow \text{count}[i]; \text{index} \leftarrow i \]
10. \[ \text{for } j \leftarrow 1 \text{ to } n \text{ do} \]
11. \[ \text{if } M[\text{index}, j] = 1 \text{ then} \]
12. \[ X[j] \leftarrow 0; P_c \leftarrow P_c \cup \{j\} \]
13. \[ \text{for } j \leftarrow 1 \text{ to } n \text{ do} \]
14. \[ \text{if } X[j] \neq 0 \text{ then} \]
15. \[ P_{k+1} \leftarrow P_{k+1} \cup \{j\} \]
16. return \( P \)
We discard the place ID’s that are not designated for consideration or already chosen as the central node for a cluster in Line 4. We count the number of nearest neighbors (distance to the neighbor is less than 1km) for each of the place ID’s under consideration in Lines 5 through 9. At the end of the loop we have an identified central node for coalescing the nearby visited places into a new cluster.

We build the cluster around the newly identified central node in Lines 10 through 12. We collect and return all the clusters constructed by the algorithm in Lines 13 through 16.

We can vary the trajectory of our greedy algorithm by supplying different seed vectors. For the default case, the seed vector consists of all one’s indicating that all place IDs are candidates to be a central node for a cluster. By making the corresponding entries for place IDs that are outside of the top 10 most visited places to be zero, we build clusters around the top 10 most visited places.

Hierarchical clustering is a bottom-up algorithm that builds a multilevel hierarchy of clusters. Given a list of pairwise distances for all visited places, \( k \), in a user’s place-to-place transitions (\( k \times (k - 1)/2 \) of distances), that links the closest pair into a cluster (each cluster contains two other clusters), leaving \( k - 1 \) clusters; we take the two closest clusters and group them. We repeat until we hit the ceiling and only have one cluster. Thus when we go down the dendogram, we observe the make-up of top clusters. We should have \( k \) levels (\( k \) different sets of clusters) of this cluster in addition to \( k \) central points.

We use the hierarchical clustering algorithm from the Statistics and Machine Learning Toolbox of MATLAB. The details of the algorithm and implementation are described in http://www.mathworks.com/help/stats/hierarchical-clustering.html.

### 5.4.2 Results

We present the aggregated average accuracies from the cases of baseline (no annotations), fifteen most frequently visited places, clusters through the greedy algorithm, and the hierarchical clustering algorithm. The averages are summarized in Table 5.4.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Agg. Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Baseline (place-to-place traces and no annotations)</td>
<td>50.47%</td>
</tr>
<tr>
<td>2  Sixteen clusters using the greedy algorithm</td>
<td>65.36%</td>
</tr>
<tr>
<td>3  Sixteen clusters using the hierarchical clustering algorithm</td>
<td>79.47%</td>
</tr>
<tr>
<td>4  Clusters based on fifteen most visited places and the incidental place</td>
<td>64.09%</td>
</tr>
</tbody>
</table>

Table 5.4. Aggregated Prediction Accuracy Averages Clustering Visited Places Based on Distance Matrices
As noted in [49], the number of user visited places ranges from 15 to 195, and the average number of places visited by a participant in the data collection campaign is 71. From Table 5.2, the baseline average considering all visited place and no place semantics annotation is 50.47%. We gain only 8% improvement in accuracy by restricting the number of visited places to the top five most frequently visited places. PeriodicaS barely gains two percentage points in accuracy improvement when the top fifteen frequently visited places are considered for next place predictions. Therefore we choose to study the next place predictions of PeriodicaS for the cases of 16 clusters.

We achieve significant improvements in accuracy of next place predictions, 28% to 58% with different clustering approaches through clustering visited places based on inherent semantics from distance matrices. PeriodicaS produces the most accurate next place predictions when using clusters determined by the hierarchical clustering method.

Feature set 2, in Table 5.4, uses no seeds for clustering visited places, while feature set 4 uses the fifteen most visited places as the seeds for clustering. Inherent semantics based on proximity of places enable PeriodicaS to make more accurate next place predictions as compared to the inherent place semantics by visit frequency or visit duration. Combining these two inherent place semantics yields mixed results as shown with feature set 4 in Table 5.4. Note that periodicity is the ubiquitous property of the place-to-place transitions and is recognized and incorporated in the next place predictions produced by PeriodicaS.

We present the aggregated average accuracy profiles for all users using the greedy algorithm in Figure 5.5 and using the hierarchical clustering algorithm in Figure 5.6. These profiles correspond to the results associated with feature sets 2 and 3, respectively, in Table 5.4. For all participants, accuracy of next place predictions improves for all users across the board and the gain in performance improvement for a given user is similar.

A typical user visits several different nearby restaurants on weekdays for lunch. These places are all represented as distinct place IDs in a user’s place-to-place transitions. With our deduced spatial inherent (place) semantics, we coalesce all these lunch places into a single cluster. Our periodicity mining techniques, published in [43] and [49], recognize and mine patterns of movement even with variations in start, end, and duration of lunch times in a user’s place-to-place transitions. PeriodicaS achieves the best results by combining the periodicity mining techniques and deduced inherent place semantics and produces next place predictions that are accurate 80% of the time across all users. This is a remarkable improvement compared to the best of the class next place predictors’ accuracy which is approximately 50%.

5.5 Multiple Choices for Next Place Predictions

A next place predictor with highly accurate predictions enables the corresponding forecasting system to produce reliable user mobility forecasts. In turn, these forecasts yield realistic and beneficial mobility models that could form the basis for exploring techniques and methods for providing a variety of communication services.

Starting with basic place-to-place transitions from a user’s mobility traces, PeriodicaS makes next place predictions utilizing periodicity mining techniques and deduces inherent place semantic annotation that are accurate 80% of the time across all users.
Fig. 5.5. Aggregated Average Accuracies for all Users with Sixteen Clusters Using Greedy Algorithm; Aggregated Average Accuracy Across all Users is 65.36%.

Fig. 5.6. Aggregated Averages for all Users with Sixteen Clusters Using Hierarchical Clustering Algorithm; Aggregated Average Accuracy Across all Users is 79.47%.
in the data set. In this section we explore ways to improve the accuracy by considering multiple next place predictions.

We use the multi-class classification mechanism of Support Vector Machines in PeriodicaS, our next place predictor. LIBSVM \cite{Libsvm} constructs a multi-class classifier by combining several binary classifiers. Using the \texttt{−b −1} option with \texttt{svmpredict}, we get the decision values associated with all place IDs as possible next place destinations. By processing these returned values from \texttt{svmpredict} further, we generate first choice, second choice, and third choice next place predictions.

### 5.5.1 First, Second, and Third Choice Predictions for Next Place

By default, the \texttt{svmpredict} routine of LIBSVM returns a single next place prediction. By turning on the \texttt{−b −1} option, we obtain the decision values generated for all place ID pairs comparing them as candidates for next place prediction. We process these decision values or probability estimates using Algorithm 3 to generate multiple next place predictions.

The input for our algorithm is the \texttt{training-label-vector}, which was returned by the \texttt{svmtrain} function of the LIBSVM. We invoke \texttt{svmpredict} with the trained model, the context of the current place and libsvm-options. We only show the training label (place) vector in Line 1 for simplicity.

**Algorithm 3:** Three Predictions for Next Place

| Input: training − label − vector |
| Output: $n_1, n_2, n_3$: first, second and third choice predictions |
| 1 $d \leftarrow \texttt{svmpredict}(\texttt{training}\_\texttt{label}\_\texttt{vector})$ |
| 2 $v \leftarrow \{0, 0, \ldots, 0\}$ |
| 3 for $i \leftarrow 1$ to $k - 1$ do |
| 4 for $j \leftarrow i + 1$ to $k$ do |
| 5 if $d[i, j] > 0$ then |
| 6 $v[i] \leftarrow v[i] + 1$ |
| 7 else |
| 8 $v[j] \leftarrow v[j] + 1$ |
| 9 return $\{p_1, p_2, p_3 | v[p_1] \geq v[p_2] \geq v[p_3] \geq v[x], x \neq p_1, p_2, p_3\}$ |

For a given number of place IDs $k$, vector $d$ consists of $k(k - 1)/2$ decision values corresponding to probability estimates computed using $k(k - 1)/2$ binary-class SVMs. Line 1 shows the assignment of these decision values returned from \texttt{svmpredict} to $d$. For example with five place IDs, vector $d$ consists of probability estimates for all five place IDs in the following order $\{(1,2), (1,3), (1,4), (1,5), (2,3), (2,4), (2,5), (3,4), (3,5), (4,5)\}$.
Line 5 tests $d[i, j]$ for a positive value; if that is the case the probability estimate is for place $i$ to be the next place prediction; otherwise the probability estimate is for place $j$. We count through majority voting of all probability estimates for one place ID at a time (Lines 4 through 8) and repeat the process for all place IDs (Line 3).

We return the three place IDs with the highest accumulated probability estimates in Line 9. The first choice for the next place prediction computed through this approach is always the same as the next place prediction returned by `svmpredict`.

### 5.5.2 Prediction Accuracy Averages

Along with the Baseline use case, we have chosen four different use cases to study the performance of multiple next place predictions and summarized the results in Table 5.5. Note that the aggregated average accuracies with one choice are identical as the corresponding accuracy averages presented in the previous sections.

The second column in the table shows the improvement in prediction accuracy when all three choices are considered for next place prediction. With multiple choices for next place predictions, accuracy improves in every use case. With three possible predictions for the next place, and using deduced inherent place semantics through hierarchical clustering, PeriodicaS makes next place predictions that are accurate 94.36% of the time across all users in the Nokia MDC data set.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>One Choice</th>
<th>Three Choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Baseline (all places and semantic annotations)</td>
<td>50.47%</td>
<td>63.97%</td>
</tr>
<tr>
<td>2 With fifteen most visited places</td>
<td>51.27%</td>
<td>73.30%</td>
</tr>
<tr>
<td>3 With five most visited places</td>
<td>54.60%</td>
<td>85.09%</td>
</tr>
<tr>
<td>4 Sixteen clusters using the greedy algorithm</td>
<td>65.36%</td>
<td>87.08%</td>
</tr>
<tr>
<td>5 Sixteen clusters using hierarchical clustering algorithm</td>
<td>79.47%</td>
<td>94.36%</td>
</tr>
</tbody>
</table>

Table 5.5. Aggregated Averages with Multiple Predictions

With a single choice for the next place prediction, as discussed before, the aggregated average accuracy improves only by 1.5%. With three choices for next place predictions accuracy jumps by 14.58%.

We present the prediction accuracy average profiles of all users with three next place predictions for the case of the fifteen most dominant visited places in Figure 5.7 and for the case of sixteen clusters using hierarchical clustering in Figure 5.8.
Fig. 5.7. User Average Accuracies with Three Predictions for Each of the Next Place Prediction - 15 Most Dominant Places; All Prediction Ave 73.30%, First Choice Prediction Average 50.85%, Second Choice Prediction Average 13.61%, Third Choice Prediction Average 8.83%
The profile in Figure 5.7 shows how each of the first, second and third choices compare against the ground truth for next place prediction, as well as the cumulative prediction accuracy i.e., next place prediction is considered accurate if the ground truth matches with any of the three choices. The first choice compares better than the second choice, and the second choice compares better with the ground truth for the next place prediction. All three next place prediction choices for a given user are consistent.

Comparing three choices for each next place prediction with sixteen clusters of all visited places across all users in the data set yields an aggregated average accuracy of 94.36%. For many users first and second choice predictions add up to or exceed 90% accuracy for next place predictions.

5.6 Individual User Prediction Accuracies

Aggregated averages characterize our ability to predict group mobility patterns. We will explore individual users’ next place prediction accuracies and their average accuracies. Individual averages are greatly impacted by a) users’ participation in the data collection campaign b) noise in the trace data collected due to poor reception, upload errors and faulty gps coordinates in recording place-to-place transitions, c) users’ life style changes during the data collection period, as described in [49], and d) the amount of periodicity in user’s mobility movements.

5.6.1 User Participation

We observe different levels of user participation during the data collection campaign. Periodicity in daily movements varies from user to user. Inherent semantics have different levels of impact on individual prediction accuracy averages.

We present the characteristics of participants’ mobility traces in the Table 5.6. The number of transitions per day, per weekday, and per weekend day, are based on each participant’s corresponding average. Traced time in hours is all the time accounted for in the user’s traces, and available trace time is calculated from a user’s sign-up to his/her exiting the data collection campaign. The same reasoning applies to entries on days participated vs total number of days.

<table>
<thead>
<tr>
<th>Description</th>
<th>Average</th>
<th>Median</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 total # of traces</td>
<td>736</td>
<td>700</td>
<td>366</td>
</tr>
<tr>
<td>2 # transitions per day</td>
<td>2.96</td>
<td>2.85</td>
<td>0.63</td>
</tr>
<tr>
<td>3 total traced time in hours</td>
<td>3814</td>
<td>3792</td>
<td>1816</td>
</tr>
<tr>
<td>4 available trace time in hours</td>
<td>9036</td>
<td>8953</td>
<td>2950</td>
</tr>
<tr>
<td>5 transitions per weekday</td>
<td>3.09</td>
<td>3.0</td>
<td>0.70</td>
</tr>
<tr>
<td>6 transitions per weekend day</td>
<td>2.52</td>
<td>2.52</td>
<td>0.58</td>
</tr>
<tr>
<td>7 days participated</td>
<td>245</td>
<td>245</td>
<td>103</td>
</tr>
<tr>
<td>8 total number of days</td>
<td>303</td>
<td>303</td>
<td>118</td>
</tr>
</tbody>
</table>

Table 5.6. Statistics from Participants’ Mobility Traces
Fig. 5.8. User Average Accuracies with Three Predictions for Each of the Next Place Prediction - Using Hierarchical Clustering Algorithm: All Prediction Average 94.36%, First Choice Prediction Average 79.47%, Second Choice Prediction Average 11.55%, Third Choice Prediction Average 3.02%.
Group participation data in Table 5.6 is an indicator of the scope and success of the data collection campaign in collecting trace data from divergent set of participants during the collection period of 10 months. There are variances in the collected traces of different users. Users’ periodic behavior is at different levels within the group. These variables contribute to making accurate next place predictions in PeriodicaS.

### 5.6.2 Selected Feature Sets and Specific Users

Next, we selected a sample of users for analyzing individual prediction averages. We analyze the individual prediction averages computed with different sets of feature vectors.

We selected 16 users and compare their prediction accuracy averages across different feature sets and against the aggregated prediction accuracy average. The results are shown in Figure 5.9.

- Individual prediction averages do not fall in a narrow band as is the case with the aggregated prediction accuracy averages.
- Some users (#64) with very low base line average do better than average with the Duration 3 feature set, and the opposite holds true with other users (#105).
- For some users (#62) prediction accuracies with one feature set (Duration 3) are much higher compared to the prediction accuracies with other feature sets.

Comparisons presented in Figure 5.9 will be helpful for $P_{b}MFS$ in determining the right feature sets to use based on the context of a specific query. $P_{b}MFS$ maintains a database of deduced inherent semantic annotations and corresponding measurement matrices for aggregated prediction accuracies as well as individual users’ next place prediction accuracy averages. Aggregated average accuracies are most useful in selecting the initial future sets at the start of a forecast generation while individual averages are beneficial in switching from a starting feature set to a new feature set to maintain the high accuracy of next place predictions. We will illustrate this aspect of differing feature sets in generating an end-to-end mobile trajectory for users in the next section.

We examine the average accuracies of next place predictions for user #47 and user #51. For this scenario we consider the users’ mobility traces with the top 15 most visited places and the incidental place; the averages are listed in Table 5.7.

<table>
<thead>
<tr>
<th>Users</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>All 3 Choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Users</td>
<td>51.27%</td>
<td>13.61%</td>
<td>8.33%</td>
<td>73.30%</td>
</tr>
<tr>
<td>User #47</td>
<td>72.61%</td>
<td>14.28%</td>
<td>8.33%</td>
<td>95.22%</td>
</tr>
<tr>
<td>Users #51</td>
<td>24.03%</td>
<td>11.53%</td>
<td>1.92%</td>
<td>37.48%</td>
</tr>
</tbody>
</table>

Table 5.7. Comparing Accuracy Averages of All Users, User #47, and User #51 with 15 Most Frequently Visited Places and Multiple Choices for Next Place Predictions
The number of traces for user #47 in the data set is 475 traces as compared to the average number of traces (736) across all users as shown in Table 5.6. The average prediction accuracies of next place prediction for user #47 are: a) 72.61% with first choice prediction, b) 14.28% with second choice prediction, c) 8.33% with third choice prediction, and d) 95.22% when all three prediction choices are compared to the ground truth for next place.

Examining the mobility traces of user #47 further, we observe a) traces are evenly distributed across the total number of days available for tracing, b) he/she never visited a place only once or twice, c) four place IDs (1, 2, 4, and 16) are dominant in the mobility traces, d) no inconsistencies in frequently visited places arise throughout the trace period. All these factors have an impact on the semantic annotations we deduce and enable PeriodicaS to make accurate predictions.

User #51 contains 175 entries, which is well below the average number of traces of all participants. The average prediction accuracies of next place prediction for user #51 are: a) 24.03% with first choice prediction, b) 11.53% with second choice prediction, c) 1.92% with third choice prediction, and d) 37.48% with all three prediction choices when compared to the ground truth for the next place.

The low accuracy average of user #51 is attributed to a) traces change in their temporal parameters and frequency through the course of the trace period, b) his/her visiting many places only once, twice, or few times c) the existence of only one dominant place in the mobility traces by visit frequency, and d) the fact that places appear and then disappear for long stretches of collected traces. These factors make many “minor classes” in the model, and next place predictions from these classes have low accuracy [43].

Through individual predication accuracy analysis, we find that many factors play a role in next place predications. Our concept of inherent place semantic annotations is applicable uniformly across all users’ place-to-place transitions. Individual user and aggregated prediction accuracy averages guide us in selecting and switching between feature vectors corresponding to different annotations in making user mobility forecasts. With insights from individual user mobility traces (swing shift work schedule or seasonal mobility of a student from semester to semester), we can adopt the definition and usage of inherent semantics for greater accuracy in next place predictions.

5.6.3 Prediction Accuracies with Temporal and Spatial Annotations

The aggregated average prediction accuracies with thirteen feature sets described in the previous sections are plotted in Figure 5.10. We selected ten participants from the data set and present their individual prediction accuracy averages. Averages for baseline, using no semantic annotations, are presented as one of the feature sets, and the baseline averages are clearly marked in Figure [5.10]. For comparison purposes aggregated average accuracy across all users for all the feature sets is plotted as a single line. Feature sets are plotted on the x-axis from low aggregated average accuracy to high.

We selected feature sets based on deduced semantic annotations for weekends, place-ID-only, visited-place-duration, weekdays and so on. While the performance of
Fig. 5.9. Individual Prediction Averages across Five Different Feature Sets

Fig. 5.10. Individual User Prediction Accuracy Averages
PeriodicaS with most of the feature sets is better than users’ baseline averages, performance with few feature sets (for example weekend-only feature set) is lower than the corresponding user’s baseline average.

We selected users with higher prediction accuracy averages, close to the aggregated prediction accuracy average, and lower prediction accuracy averages. The performance of PeriodicaS is consistent, from feature set to feature set, for users with higher periodicity. Prediction accuracies for users with lower periodicity (user #12, user #87, user #103, and user #142), particularly feature sets that perform lower than the baseline, jump up and down compared to their corresponding baseline averages.

Next place predictions based on some feature sets (for example weekend based annotations) have lower accuracy averages than baseline averages. However, some individuals (user #12) with periodic behavior in their weekend movements (WeeS) through their work-place arrangements and other lifestyle related mobility movements have higher next place prediction average accuracies due to these factors.

5.7 Mobility Forecasting

$P_{b}MFS$ - Periodicity based Mobility Forecasting System - projects user’s movements, forecasts end-to-end user trajectories, and provides answers to a variety of queries related to users’ movements. Next place predictions are the atomic units of these forecasts and PeriodicaS is the core component of $P_{b}MFS$.

The four major components in $P_{b}MFS$ are: a) a data set with place-to-place transitions from a diverse group of participants, b) feature vectors built on periodicity in user movements and deduced semantic annotations, c) the next place prediction algorithm PeriodicaS, and d) precomputed prediction accuracy averages as measurement matrices.

5.7.1 Operational Flow of $P_{b}MFS$

$P_{b}MFS$ systematically follows the steps in the Knowledge Discovery in Databases (KDD) Process for generating user mobility forecasts. We describe the operational flow of $P_{b}MFS$ below.

1. $P_{b}MFS$ starts with a user’s place-to-place transitions in their mobility trace files. Traces for each user are used for the purposes of training and building the measurement matrices.

2. The system compiles sets of available place annotations and deduces inherent semantic annotations. $P_{b}MFS$ constructs feature sets corresponding to these annotations to be used by the next place predictor, PeriodicaS.

3. Based on the type and context of a query, $P_{b}MFS$ determines the feature sets that are the most suitable for next place predictions using the database of measurement matrices.

4. $P_{b}MFS$ invokes PeriodicaS several times with the appropriate feature sets to build a forecast corresponding to the query.
In the following section, we present two examples of queries and resulting forecasts. We demonstrate the flow of $P_b MFS$ through the examples.

$P_b MFS$ handles a variety of queries and generates forecasts for individual users’ future mobility. In this section, we illustrate the applicability of $P_b MFS$ through two use cases: a) generating an end-to-end trajectory for a given user, current place, current time, and duration in hours, and b) generating a yes or no forecast whether a given user, starting at a time-stamp ($t_{cur}$) will move from a place of interest ($p_{cur}$) to another place of interest ($p_t$) within the specified time duration.

5.7.2 End-to-end trajectories

We forecast users’ end-to-end trajectories and compare them against the ground truth. We are not predicting multiple place-to-place (disjointed) transitions, but building trajectories with successive next place predictions.

We examine the test sequence for users’ place-to-place transitions with a given start time, sometimes spanning across multiple days. Some of the end-to-end trajectories in our data set starting at the most dominant place (user’s home) and ending at the same place result in short predictable trajectories. We utilize predictions generated for the trace data set with the top ten frequently visited places and the incidental place (with ID 11). We present the end-to-end trajectories for users #14 and #26 in Figure 5.11 and Figure 5.12.

From the results from PeriodicaS, the aggregated average accuracy of next place predictions across all users with the feature set consisting of the ten most dominant places is 54.68%. The average next place prediction accuracy for user #14 is 64.29% and for user #26 is 58.23%.

After two transitions, user #14 moves from place one to place four. We predict place 11 (incidental place) as the next destination place from place one compared to the ground truth of place four. Then the forecasting system recovers and predicts place one accurately as the (returning) destination place.

An example of end to end trajectories, actual and predicted, of user #26 are shown in Figure 5.12. From place ID one both the prediction and the ground truth are identical, composite place ID 11. We represent the composite place ID 11 in the end-to-end trajectory as two different nodes, as the composite ID contains different places. Place IDs three and one are represented as different instances, indicating different time-stamps of the user’s visits. Feature sets are switched at time-slots A and B in Figure 5.12. In this example the projected and actual end-to-end trajectories are identical.

5.7.3 Forecasts for Places of Interest

Here we are interested in a user’s movements from one place of interest ($p_{cur}$) to another place of interest ($p_t$). In this case generating mobility forecasts based on user traces consisting of all visited places will be inefficient. We make use of our trace filtering techniques based on inherent semantics.

First we generate a feature set with $p_{cur}$ and $p_t$, and all other places visited by the user are combined into a composite place ID known as incidental place. We preprocess the user mobility traces appropriately using this feature set for PeriodicaS.
Fig. 5.11. End-to-end Trajectory Example for User #14

Fig. 5.12. End-to-end Trajectory Example for User #26 with Switching of Feature Vectors at A and B
During successive invocations of PeriodicaS, $p_{\text{cur}}$ is also incorporated into the *incidental place*. We show the available next place choices in the initial and subsequent invocations of PeriodicaS in Figure [5.13].

As many places are combined in the *incidental place*, there could be place-to-place transitions from the *incidental place* to itself. In this scenario, with only two places of interest and the rest combined into an *incidental place*, the average next place prediction accuracy will be around 44.4% for an algorithm that makes predictions randomly.

Fig. 5.13. Next Place Predictions Choices with Current Place, Target Place, and Incidental Place

---

**Algorithm 4: User Mobility Involving Places of Interest**

- **Input**: current place $p_{\text{cur}}$, current time $t_{\text{cur}}$, number of hops $n$, target place $p_{t}$, and $\{\text{inc\_place}\}$
- **Output**: *yes* if the user will move from $p_{\text{cur}}$ to $p_{t}$ in $n$ hops or less, or *no* if the user will not visit $p_{t}$ during the $n$ hops

1. $p_{\text{next}} \leftarrow \text{PeriodicaS}(p_{\text{cur}}, t_{\text{cur}})$
2. if $p_{t} = p_{\text{next}}$ then
   3. return *yes*
4. $\{\text{inc\_place}\} \leftarrow p_{\text{cur}}$
5. $p_{\text{cur}} \leftarrow \{\text{inc\_place}\} + \{t\}$
6. for $\text{ind} \leftarrow 2$ to $n$ do
   7. $t_{\text{cur}} \leftarrow t_{\text{cur}} + \text{hop\_dur}$
   8. $p_{\text{next}} \leftarrow \text{PeriodicaS}(p_{\text{cur}}, t_{\text{cur}})$
   9. if $p_{t} = p_{\text{next}}$ then
      10. return *yes*
11. return *no*
$P_bMFS$ converts the time duration into a number of hops ($n$), and invokes PeriodicaS successively with the appropriate feature set. We describe the algorithm used by $P_bMFS$, Algorithm 4 below.

Lines 1-3 make the initial next place prediction using PeriodicaS. If a user is not expected to move to $p_t$ through the initial next place prediction, then the current place, start time, and end time are adjusted for the next iteration of invoking PeriodicaS. The average duration period is precomputed by $P_bMFS$ from the place-to-place transitions in the user’s trace file.

When $p_t$ is the predicted next place during one of the $n$ trials, $P_bMFS$ will terminate the process and forecast that the user will visit place $p_t$ during the given time duration. If $p_t$ is not the predicted next place in any of the $n$ tries, then $P_bMFS$ forecasts that the user will not visit place $p_t$ within the given time duration.

The aggregated next place prediction accuracy average with two places is 71.35% and with one place is 81.77% as shown in Figure 5.3 and Figure 5.4 respectively. $P_bMFS$ makes the forecast with respect to user movement to $p_t$ with higher accuracy than the Baseline average of 50.47%. This prediction accuracy applies to a negative forecast when $P_bMFS$ determines that the user will not move to $p_t$ in the given time duration.

5.7.4 Trajectories with Multiple Choices for Next Place Predictions

We generate end-to-end trajectories for two users (user #58 and user #66) with three choices for next place predictions. There are three levels or three hops in the forecasts, and they are shown in Figures 5.14 and 5.15. We shade the place ID’s that match the ground truth.
Fig. 5.14. Forecast for User #58 with Three Choices for Next Place Predictions at Each Hop; Next Place Predictions Matching with the Ground Truth are Shown in Gray

For these predictions we use baseline (no semantic annotations) traces. The next place prediction average accuracy for user #58 with three choices for next place predictions is 71.97%. For the very first next place prediction in the forecast, the second choice matches with the ground truth. For rest of the forecast, the first choice of the next place prediction matches with the ground truth.
We use the greedy algorithm with sixteen clusters for generating the forecast for User#66. The next place prediction average accuracy for user#66 with three choices for next place predictions is 91.57%. In the generated forecast the first choice for each of the predicted next place matches with the ground truth.
5.8 Summary and Discussion

$P_bMFS$ is a versatile system that forecasts a user’s future place-to-place movements based on their available mobility traces. We coopted $PeriodicaS$, our proven next place predictor, as the core component of $P_bMFS$. Periodicity and place semantic annotations in users’ mobility traces are leveraged by $PeriodicaS$ for accurately predicting next place destinations. We deduce inherent semantic annotations and convert these annotations into a rich set of feature vectors.

We start with mobility traces consisting of anonymized place IDs and timestamps. By mining periodicity in user movements, $PeriodicaS$ predicts next place destinations given a future place ID and a time stamp. Place semantic information and temporal information will yield more accurate predictions [49]. When place semantic information such as home, eatery, and work place etc., and user demographics are available, they improve prediction accuracies by 54%.

Annotations that are an integral part of the trace data set are preferred compared to the annotations compiled through user forms. We observe that the annotations which come through forms or other mechanisms are sparse, and accuracy can not be independently verified. We expand the concept of inherent semantics from [49] and show in some instances that these annotations improve prediction accuracies by 62%.

$P_bMFS$ utilizes information available in the data sets as well as deduced annotations to drive $PeriodicaS$ to generate better next place predictions. Deduced inherent semantics are user-specific and are not shared between users. While generating a single mobility forecast, $P_bMFS$ dynamically adjusts feature vectors supplied to $PeriodicaS$ based on the context of successive predictions.

We give two applications of $P_bMFS$ - generating forecasts of end-to-end trajectories, and yes or no mobility statements involving places of interest. We compared the generated forecasts with the ground truth, and the generated forecasts are reliable representations of future in terms of mobility.

Mobility forecasting plays an important role in digital assistants [50] in providing just-in-time context aware services. $P_bMFS$ is suitable for deploying in digital assistants. In these environments future predictions and the corresponding results become part of the training trace data. This will enhance the continuous learning and the flexibility of $P_bMFS$. Future mobility patterns generated by systems such as $P_bMFS$ could be utilized by the digital assistants to perform autonomously for the benefit of the user [24].

Mobile traces that are accurate and less noisy improve the efficiency of next place predictors. Trace data collection can be improved by having the users involved actively. Issues related to the aging of traces [49] and assigning proportional weights to different traces can be resolved with users’ feedback and configuration.

We explored combinations of different feature sets, weekdays and top three frequently visited places, and for different users different combinations yield better forecasts. We employed standard methods in switching feature sets from prediction to prediction in generating longer forecasts. This mechanism could be automated and improved to generate better forecasts.
Chapter 6

Models for Opportunistic Networks and Moving Forward

We embarked on research activities related to predicting users’ next place destinations and users’ mobility forecasting based on their mobility traces to develop realistic models for opportunistic networks. Opportunistic networks [51] are an interesting evolution of MANETs, mobile ad hoc networks. Mobile nodes are enabled to provide or receive communication services with each other or the access-point servers even if a route connecting them never exists. Instantiations of opportunistic networks are going through fast paced changes and are becoming ubiquitous through Smart City scenarios [52].

Reliable models of opportunistic networks are needed for developing, and fine tuning, efficient service deployment strategies. Human mobility traces provide an excellent basis for generating models of opportunistic networks [33].

Features based on users’ mobility forecasts will enhance the functionality of digital assistants, such as SIRI and Cortana, significantly [24] and [25]. Finally, we will examine use-cases where the effectiveness of digital assistants improves with a better understanding of users’ future mobility patterns.

6.1 Opportunistic Networks

Hui and Crowcroft [11] propose an opportunistic communication system design based on human mobility models. Two types of models for opportunistic networks can be generated from our end-to-end trajectories of $P_{b,MFS}$: a) Individual User Oppnets and b) Group Oppnets.

When a user is at home, her smart phone is part of a network of server(s) and other smart devices in the home as shown in Figure 6.1. As the user transitions from place-to-place during the day her smart phone plugs into different networks. Based on the user’s mobility forecast generated from $P_{b,MFS}$, we can model this user’s oppnet for the day.

With data sets consisting of mobility traces of a group, we can determine the meeting-points between users. We can also predict future group oppnets for a set of given users by hop-count. $P_{b,MFS}$ will generate a backup prediction for user-plus-location combination where the prediction accuracy is low. We generated reliable models for group oppnets with first and second choice predictions from $P_{b,MFS}$.

The MDC data set is most suited for generating individual user oppnets. With many anonymization techniques [30] employed during the data collection campaign, common place-IDs are not deducible between different users. The Wireless Topology Discovery data set [27] contains the Access Point information as part of the students’ mobility traces. With this shared Access Point Server information, we can generate reliable models for group oppnets. We select students with common attributes such as belonging to
Fig. 6.1. User’s opportunistic network connections through the day: at home, work, lunch and other visited places

the same residence hall or being registered for the same course and utilize the mobility forecasts generated by $P_{MFS}$ for the selected group of students.

6.2 Summary and Conclusions

Mobility networks built from users’ place-to-place transitions are complex. We utilize techniques and methods that recognize and exploit periodicity in human mobility as the basis for pattern mining of these raw user mobility networks. Based on place-to-place transitions collected over a long period of time (four to ten months) with a diverse group of participants, we built individual user mobility models and predicted their next place destinations.

Next place prediction is a powerful tool in determining the users’ location, which is essential for data delivery and location services. Predicting two or three potential next locations is also useful; however, having accurate next place prediction is important for communication services such as data forwarding and replication, and store-carry-and-forward data dissemination strategies. We quantify the accuracies of our predictions, as they provide guidance in selecting a suitable feature set based on the query for next place prediction.

We characterize users’ mobility patterns through a systematic study of their place-to-place transitions, and we predict users’ next place destinations. We compute individual prediction accuracies by averaging several next place predictions for the corresponding user, and in turn, based on the individual prediction accuracies we generate an aggregated prediction average accuracy for the group. We use this measure to characterize our ability to mine user periodicity in collected place-to-place transitions through our next place predictor, PeriodicaS, as well as gauge the benefit of semantic annotations.

We deduce inherent semantics annotations from the place-to-place transitions and use these annotations to derive new feature sets for PeriodicaS. We show that these
annotations, along with place semantics and demographic information available in the data set, improve aggregated next place prediction accuracies.

6.3 Future Work

Our next place predictor, PeriodicaS, was noted for its elegance and simplicity in mining available periodicity among the submissions for the Nokia’s Mobile Data Challenge. In comparison, Etter at all [44] use a combination of dynamical Bayesian Networks, artificial Neural Networks and gradient boosted Decision Trees as part of their next place predictor to achieve a gain of three percentage points compared to the aggregated average prediction accuracy of PeriodicaS. The core algorithms of our next place predictor can be improved by experimenting more with SVM RBF (Radius Basis Function) kernel parameters, using different SVM kernels, and augmenting with other data mining techniques. Translating a three percentage point gain from baseline averages to the highest averages we achieved, 94% aggregated accuracy average across all users with annotations derived from the hierarchical clustering algorithm, would be significant.

Many other semantic annotations can be deduced based on the examples provided in Chapter 5, as more useful annotations give added flexibility to $P_bMFS$. We can supply weights to the features deduced from semantic annotation; this mechanism gives us added flexibility in $P_bMFS$ for generating better forecasts compared to selecting or deselecting feature sets.

Mobility forecasts generated by $P_bMFS$ improve the efficiency of use-cases in digital assistants. For example, Bump is an iOS and Android application that enabled smartphone users to transfer contact information, photos and files between devices (https://en.wikipedia.org/wiki/Bump_(application)). A user’s mobility forecasts would enhance this feature by scheduling Bumps in the future.
Bibliography


[6] Nadia Drake. Milky way has 4 billion years to live - but our sun will survive. *National Geographic*.


[29] Zhenhui Li, Jingjing Wang, and Jiawei Han. Mining event periodicity from incomplete observations. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD ’12, pages 444–452, New York, NY, USA, 2012. ACM.


[41] Zhenhui Li, Bolin Ding, Jiawei Han, Roland Kays, and Peter Nye. Mining periodic behaviors for moving objects. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD ’10*, pages 1099–1108, New York, NY, USA, 2010. ACM.


[43] Bhaskar Prabhala, Jingjing Wang, Budhaditya Deb, Thomas La Porta, and Jiawei Han. Leveraging periodicity in human mobility for next place prediction. In *Wireless Communications and Networking Conference (WCNC), 2014 IEEE*, pages 2665–2670, April 2014.


Vita
Bhaskar Prabhala

Education

*The Pennsylvania State University*
Ph.D. in Computer Science and Engineering – December 2015
M.S In Computer Science – May 1979

*Hindu College, Guntur, India*
B.Sc. in Mathematics and Physics – May 1972

Experience

*Philips Semiconductors / Royal Philips* 2001–2009
General Manager, Development Center – San Jose, CA & Shanghai, China

*Real Communities Incorporated, Cupertino, CA* 1998–2001
Co-founder and Vice President of Product Development

Vice President of Engineering