A Data Mining Approach To Rapidly Learning Traveler Activity Patterns For Mobile Applications

BY

CHAD A. WILLIAMS
B.S., Cornell University 1998
M.S., DePaul University 2006

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Chicago, 2010

Chicago, Illinois

This dissertation is dedicated to three people who shared with me the sacrifices required to complete it. The first is my wife, Patricia Boye-Williams, who shared in my struggles of trying to be a full-time student, husband and father. Without her emotional support and encouragement the completion of this project would not have been possible. The other two are my daughters, Grace and Kate Boye-Williams, who are growing up into wonderful little people before my eyes. Also, I would like to thank my parents, KC and Theresa Williams, for their nurturing and support throughout my life.

ACKNOWLEDGMENTS

I am pleased to thank my co-advisors Peter C. Nelson and Abolfazl "Kouros" Mohammadian for their continued support and suggestions in the preparation of this work. Their guidance in navigating the delicate balance of an interdisciplinary dissertation was sincerely appreciated. I would also like to thank my committee members Bing Liu, Tanya Berger-Wolf, Sean Doherty, and John Polak for their valuable feedback in improving this dissertation. I would also like to thank Joshua Auld whose help related to the UTRACS survey was instrumental to the success of this research, and Martina Z. Frignani's efforts in conducting the survey. This research was funded by the National Science Foundation IGERT program under Grant DGE-0549489.

CAW

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (Continued)

FIGURE **PAGE**

LIST OF ABBREVIATIONS

TAZ	Transportation analysis zones
TASHA	Toronto Area Scheduling Model for Household Agents
ANN	Artificial neural network
GPS	Global positioning system
EM	Expectation maximization
MAR	Missing at random
ML	Maximum likelihood
MVC	Missing Value Completion algorithm (Ragel and Crmilleux, 1999)
kNN	k-Nearest Neighbor
PDA	Personal data assistant

SUMMARY

The swift growth in the number of GPS devices has led to a boom in the number of mobile applications attempting to exploit this rapidly growing market. As a result, understanding travelers and their information needs has become a major topic of interest. While many studies have examined learning traveler behavior, they have primarily concentrated on the destination and route information. There are two key weaknesses of these studies. First, they require a lengthy history of the person be collected before a reasonable model can be built. Second, they focus on the travel itself rather than the reason for the travel. While trip information is useful, the reason for the travel likely is more useful to mobile applications aimed at influencing the user plans. The purpose of this study is to address both of these points: reducing learning time and examining the reason for the travel rather than just the trip itself.

To accomplish these goals, this work examines using an interdisciplinary approach to combine transportation planning activity-based modeling methods with data mining techniques to learn individual patterns. This work demonstrates that such a model can be tailored to the patterns of an individual traveler, allowing projections of their future trips, activities, and planning flexibility to be made. Second, due to the abstraction of the model, an extensive history of the user is not necessary to build a reasonable model of the traveler. Traditionally, however, this type of model has required collecting a detailed activity history that is likely more burdensome than most mobile application users would accept.

SUMMARY (Continued)

This research addresses this challenge by creating an activity model of a traveler while greatly reducing the data entry required by the user. The primary contribution of this work is a set of techniques for quickly learning the travel activity patterns of individuals with limited user interaction. This is achieved through three main areas: (1) leveraging passive data to augment user entered data; (2) introducing techniques to reduce the impact of missing data on prediction quality; and (3) supplementing user patterns with general patterns from other sources.

CHAPTER 1

INTRODUCTION

This dissertation is a report on a research study of learning and predicting the activity and travel behavior of individuals in an automated fashion for mobile applications. This document is organized into five chapters. This chapter provides background, motivation, the problem statement, the significance of this work, and an overview of the methodology used in this work. It concludes by presenting the delimitations of this study as well as defining terms used within this dissertation. The second chapter contains a detailed review of the relevant theoretical and experimental literature on the topic. Chapter 3 explains in detail the research methodology that was used in this study, and how it was executed. The results of this research and an analysis of the findings are presented in Chapter 4. The final chapter summarizes the outcomes of this work and discusses the implications of this thesis.

1.1 Background

Transportation planners have studied travel behavior extensively over the years. More recently, the focus has shifted from looking at travel alone to understanding why a trip was made and when this decision occurred. One of the approaches used for this has been examining the activity needs/desires of the person as the reason the travel is made. Thus the information about the activity and travel options drive the decision of where the activity will take place and how the person will choose to get there. This study proposes this type of information is

more important than travel and location information alone in understanding what information is useful to the traveler *before* they make their trip.

Modeling and predicting travel patterns has been a source of interest and a challenge for regional planners for a number of decades. Recently ubiquitous computing, has begun to develop an interest in predicting travel patterns as well. The goal of regional planners has been to understand the activity and travel patterns of individuals as a means to more accurately predict the patterns of an area or population as a whole. In contrast, the goal of the ubiquitous computing community has been to predict the travel of a specific individual with little attention paid to the reason for the travel or activity. Because of their differing goals, the two approaches have significant differences in how they model travel patterns as well.

In regional planning, the goals and methods used in examining traveler's patterns have evolved substantially over the past years (Ashiru et al., 2004; Miller and Roorda, 2003; Lee and McNally, 2003; Lee and McNally, 2006). In the more distant past, this type of prediction was used primarily at a high level for predicting trips at a transportation analysis zone (TAZ) level. This type of analysis provided good projection of the demands on the overall road system at an aggregate level, but provided little insight into the details of these trips. Questions such as when certain trips were likely to occur, or the reason for the trips remained a challenge. More recent work has tried to fill in these gaps through a behavioral analysis approach. By modeling trips as the result of the activity needs of a traveler, a model can be created that addresses the details of trips while still being general enough to serve as a model across travelers.Using this

type of model, micro-simulations have been able to produce a more fine-grained projection of the aggregate behavior and trips of a population.

Recent work has begun to examine not just the purpose of trips, but also how travelers as a whole schedule their activities to accomplish their activity demands. The purpose of the majority of these efforts has been to synthesize the travel patterns for a population of interest. While these techniques simulate the activities and travel of either individuals or households, the generated travel and activity schedules are analyzed for fit at an aggregate level. As such, the predictive value of modeling the patterns of individuals has been examined at an aggregate level rather than the accuracy of any specific individual. As a result, most models in this area are created from travel surveys given to a sample of the population in question. Historically these travel surveys were primarily paper based relying solely on the participant's ability to recall their actions for a specific day. More recent surveys, such as the one introduced in this work are beginning to move to a more on-line-based approach combined with global positioning system (GPS) data to more accurately capture the complete travel patterns of the individual. The data collected in these surveys has then been used to create models of similar individuals within the population.

Ubiquitous computing on the other hand has focused on modeling an individual traveler. These models have primarily taken the form of processing GPS traces of an individual to model future movement of that person. Specifically the models generated might predict things such as when the next trip is likely to occur, the choice of route, or even the next several stops. However, the models created in this area are very much specific to the individual and the traces

of that individual. As a result, a lengthy history of GPS traces of that individual need to be collected to ensure reasonable coverage. Another limitation of this approach is that the data collected is specific to the locations visited and cannot be used beyond the specific area that was traced. As a result, these models are not able to provide any insight if the traveler visits a previously unseen location.

This research intends to combine aspects of these two approaches, to model the **activity patterns** and travel behavior of individuals. This work appears to be the first to address the integration of these two goals. Specifically this effort models not only the travel patterns of an individual, but also the reason for travel for that individual. It is important to note that while GPS is becoming more common in cars in the form of navigation systems, the spread of GPS units that remain with the user throughout the day, such as GPS enabled personal data assistants (PDA), trail significantly behind at the time of this writing. This is important since only these devices would capture trips outside of the car such as by walking or by public transportation. Thus the purpose of this study was to examine how devices such as these can be leveraged to create a richer model of the traveler as part of a larger goal to enable mobile applications such as an intelligent traveler's assistant.

1.2 Problem statement

Considerable research in urban planning has focused on predicting likely activity and travel behavior of people based on their general characteristics. These methods forecast behavior well at an aggregate level, however they are neither designed nor intended to model the idiosyncrasies of real people's schedules or predict the behavior of a specific individual. By contrast, current

techniques for modeling individuals rely on a lengthy history of the individual to be observed before a reasonable model can be made, and even so do not provide insight as to the reasons behind the travel. The purpose of this work was to quickly learn the travel behavior and activity context of an individual traveler; as such, the predictions should be tailored to the specific patterns of that individual. It was neither expected nor was it the goal of this work to identify a model that when trained for one person applied well to other people. Instead, the focus was on the ability to quickly learn the habits of an individual and adapt to changes in these patterns over time from GPS traces and user interactions. The output of this work was evaluated based the ability to make reasonable predictions for the traveler quickly and adapt to their personal patterns as additional data was gathered. A non-trivial aspect of this problem is being able to learn a reasonable model of an individual's behavior quickly without requiring a huge data entry burden on the traveler for an extended period. While there is likely some amount of initialization necessary, the burden should reduce over time, while still adapting to changes in patterns. A second aspect to be addressed was being able to identify the factors influencing behavior for the benefit of mobile applications.

Thus the goals of this study are:

1. Minimize the amount of data required to be actively collected from the traveler
2. Significantly reduce the amount of travel history required before meaningful predictions can be made
3. Create a model of the traveler that can reliably predict aspects of travel context, and

4. Develop a prediction model that can be used by mobile applications for forecasts of activity patterns as well as being able to articulate the factors influencing these forecasts

1.3 Significance of the study

This study contributes to existing knowledge in four main areas: transportation planning, ubiquitous computing, travel pattern transferability, and data mining methods. This research contributes to the scientific community through techniques to learn individual travel behavior faster than current techniques based on the observations of a single individual alone. In addition to experimentally evaluating the problem described above; this work proposes a conceptual modeling framework for future studies in predicting behavior and individual schedule flexibility.

1.3.1 Contribution to transportation planning

This research is significant because it extends our understanding of what can be learned about the travel patterns of individuals unobtrusively. Specifically this work addresses the question of how much can be learned about a traveler's preferences, schedule constraints, activity and travel behavior without over burdening people. While a model of these traits could be learned based on extensive questioning of an individual's travel over an extended period, this work seeks to do so while minimizing the effort that must be made by the traveler. This aspect is important since if the effort required by the participant is too high, they are less likely to continue the process for any extended time. This problem is readily apparent in transportation travel surveys. Many of these studies have focused on short data collection periods because of the resistance of participants to complete the amount of questions asked over any period longer than a few days (Stopher, 1992). Thus reduction in burden is one of the critical components

to enabling long term travel surveys as well as for any other applications that wish to observe and model user behavior over any extended period (Doherty et al., 2001).

A primary goal of this work was to contribute to the field's understanding of how this same type of data may be collected over an extended period while significantly reducing the burden on participants. This study presents ways the amount of data collected can be alleviated through reduction in the number of questions that need to be asked. Other ways presented reduce the information needed to be entered by the user by making use of the information collected over an extended time and continually updating the model of the user.

Other studies of modeling traveler behavior have focused primarily on capturing activity patterns of a class of individuals or capturing the travel of individuals. On the transportation planning side techniques such as activity modeling have been used to model the behavior patterns of classes of individuals. These models can then be used for micro-simulations of an area of interest. This work contributes to these areas in two key ways.

First, the techniques in this study to reduce the burden of collecting data are shown to greatly reduce the effort of a travel survey participants. This will likely make recruiting participants easier for longer term studies since the effort required by them in the data collection process could be significantly reduced. By reducing the ongoing burden this type of benefit lends itself to enabling long term studies of activity based travel behavior which have essentially been impractical in the past. To illustrate this point this work introduces a new type of travel survey that utilizes the techniques of passive collection combined with learning intro-

duced in this work and its benefits are shown empirically in the reduction of participant burden over the course of a multiple day travel survey.

Second, this work demonstrates how to better model the patterns of an individual rather than classes of individuals. Also the models created by the methods introduced in this work are proven to generate realistic orderings of activities at a per traveler basis rather than just at an aggregate level. These contributions should help micro simulations become much more realistic at a detailed level than current approaches.

A key difference between this work and prior work was the focus on the behavior of a single individual. While planners have modeled individual behavior as well, the goal has primarily been to model the behavior of a type of individual that can be generalized across similar individuals. As a result, while the behavior in general can be modeled fairly well, idiosyncrasies and unusual preferences are not likely to be accounted for. In this work, while this type of model will likely help the application learn common behaviors more quickly, the goal was to learn the specific travel preferences and idiosyncrasies of the person using the device. While these idiosyncrasies might not translate well across travelers, this subtle shift will facilitate moving travel prediction from the realm of urban planning to travel analysis for the individual for applications such as smart traveler assistant devices.

1.3.2 Contribution to ubiquitous computing

Another area that will benefit significantly from this work is the area of ubiquitous computing. Much effort has been made within this field to learn the likelihood of travel and activity taking place, but little has been gained in terms of the reasons for the behavior. This work

contributes significantly in this area by adding onto existing techniques for modeling travel by also adding the ability to learn activities, mode preferences, time constraints, and location constraints. This additional information offers substantial benefits in terms of the context data that can be used by mobile applications. Second, being able to provide the factor(s) influencing this behavior will make it possible for applications to try and influence the traveler's behavior.

The principle behind these smart traveler's assistant devices are to provide intelligent services like multi-modal travel planning to support a user's travel needs (Dillenburg et al., 2002; Dillenburg et al., 2004; Torrens et al., 2004). With this type of application, understanding the context of a traveler can help identify what type of information the user is interested in and help narrow potential options. Consider a scenario where a user wants to make a trip across town mid-day. If the assistant knew the user's likely mode of transit and accessibility thresholds, the system might be able to evaluate and suggest one transit route over another or even provide an incentive based on the person's personal preferences to choose a particular mode or destination.

Another application of this type of knowledge, would be the ability to suggest optimizations to the person's schedule to reduce travel time (or cost) that are more likely to fit within their unique travel preferences and scheduling needs and thus more likely to be followed. Unlike programs like dynamic route choice that try to help the traveler avoid congestion once the traveler is already in route, the knowledge of the flexibility of scheduling and destination could be used to help the person best select their destinations and departure time to minimize the congestion in general, or suggest the mode(s) that will work for all of their planned activities.

The ability to model these needs and addressing how they can best be achieved for the individual is a major contribution of this work.

1.3.3 Contribution to travel behavior transferrance

Another contribution of this research is demonstrating that traveler patterns can be transferred across different locations. While learning through comparing similar patterns is not new, doing so with traveler data presents some unique challenges. Unlike most collaborative learning, the similarity of travelers is not as easy as saying one traveler is similar to another traveler. Due to the differences in the physical locations in terms of travel time to activities and modes available, it is possible to be very similar in one aspect while nearly the polar opposite for some other aspect. It is in part because of this challenge that activity data from one city has not been used in a different city in transportation planning. Thus demonstrating traveler's activity patterns can be transferred marks a significant advance in the way activity data modeling and collection can take place. Furthermore this study verifies this transferability at an individual level showing that the transferred patterns also produce similar activity schedules. This contribution is significant because of the opportunities it opens up in terms of reducing traveler specific or location specific data requirements.

Another application of this technique outside the scope of this work would be reducing the number of people required for a travel activity survey by augmenting the collected patterns with those observed in other cities. This could potentially lead to significant cost savings by planners since large-scale survey efforts are often very expensive (Richardson et al., 1996).

While this work may provide additional insight for urban planners, it also enables a new realm of personalized travel analysis that might be used for individual benefit as well.

1.3.4 Data mining

One of the most significant contributions of this work came from addressing a sub problem of this study related to the presence of missing values. This work presents a new sequential associative mining and rule technique that far out performs existing techniques for data sources that have a significant amount of missing data. This technique is shown to produce better predictions for both data sources with values missing at random (MAR) and with realistic distributions of missing values.

1.4 Objectives

In this work, we investigate learning and predicting travel and activity of individuals as part of a larger goal of enabling intelligent traveler's assistants. This work combines the analysis techniques from activity modeling and scheduling with a data mining approach to identify common activity and travel patterns for use by mobile applications. This research demonstrates that these learned patterns can be used to model an individual's travel patterns much quicker than traditional methods. As shown in this work, a prediction model based on this combination of techniques can quickly learn the travel patterns of an individual, while continuing to adapt and improve the model as additional behavior is observed. As the desire for smart traveler devices grows, prediction of future travel behavior will be a key function to further enable these devices.

In this work, we focus on techniques for extracting meaning from these new data sources, such as GPS enabled PDAs, and utilizing the derived knowledge for the benefit of the device owner. One of the challenges associated with understanding traveler behavior is its strong dependency on the transportation network. The strength of this relationship makes it nearly impossible to address behavior without also considering the overall **traveler context** in terms of location, mode of transit, what has been done recently, and the time of the activity to name a few aspects. A better understanding of how this context affects future behavior is likely to not only improve predictive models, but also provide additional insight into how choices are made among alternatives. The goal of these improved predictive models is to enable personalized scheduling and travel advice.

1.5 Overview of methodology

In this section an overview is given of the methodology applied in this study, with the purpose of helping set the context for later chapters. The sub-problems of addressing the overall goal are briefly described as well as the approach to answering these challenges. The full methodology is explained in detail in Chapter 3.

For this study, a quantitative research design is used to empirically demonstrate the findings of this work. A quantitative design was chosen since it is best suited for the task of data mining and prediction of known variables. The quality of the predictions made and the anticipated time savings of the methods introduced in this work were evaluated empirically.

Specifically the problem of given a sequence of the prior traveler context was examined with respect to how well the traveler context of the traveler's next activity can be predicted.

Rather than focusing on prediction of common activity sequences based on the history of many travelers, the paradigm is shifted to predicting the traveler context of an individual based on their own travel history and the history of others. A novel approach to predicting individual traveler behavior and associated context from a stream of their prior activities is proposed. This suggested approach examines each step in their history and the predicted next step of the traveler as a set of attributes or characteristics describing the context of that step. As a whole, these steps form a sequence of sets describing the traveler context progressing through time, or an enhanced activity sequence, referred within this work as activity patterns.

There are a number of different ways to evaluate the success of being able to learn an individual's activity patterns. Other studies have evaluated this task based on route selection, geographic proximity, timing, gross activity distribution, gross mode preference, distance from optimal, et cetera (Ashbrook and Starner, 2003; Mohammadian and Doherty, 2006; Ettema et al., 2007; Frank et al., 2008; Janssens et al., 2007). Since the goal of this study was to model the patterns of an individual in a form that would be useful for predicting the next activities at a micro level, measures that focus on summarization are insufficient. Likewise metrics that focus on geographic matching miss that while the location may be different the mode of transportation or activity might be the same. As a result the metrics used in this study were chosen to combine different aspects of these measures that focus on the order of activities, the activity characteristics, and the travel characteristics, since they are the primary interest.

To demonstrate the validity of this approach with respect to these metrics, the data for this study was activity-based travel surveys. This data set was chosen because it captures the

activities of the studied travelers as well as their movement history. Since the goal of this work was to introduce ways the activity and travel of a person can be collected and predicted, this data set is well suited to the task as both of these key traits are captured in detail. While the proposed method of learning and prediction relies on GPS data and ongoing data collection, these data sets provide a way to evaluate predictions using a much larger set of data than could be reasonably collected over the course of this study. As a result the findings in this work are based on a combination of both a collection of travel surveys conducted by other researchers and planning agencies as well as a collection effort designed specifically for this study. The way these various different sources are combined is discussed in detail in Chapter 3.

To demonstrate the soundness of this work; the primary goal was broken down into a number of sub-problems. First, the internal validity of being able to predict the activity patterns of a traveler with "perfect" information is shown. Next, techniques are introduced and analyzed for reducing the burden on the traveler for collecting data. This is followed by presenting algorithms and techniques that can further reduce the data required with limited impact on prediction quality. Afterward, a demonstration of how activity patterns might be transferred from other locations to improve predictions is shown. Finally, a combined framework of how all of these concepts might be brought together for quickly learning the activity patterns of an individual is presented.

1.6 Delimitations of the study

Demonstrating the ability to learn an individual's activity patterns quickly is the primary aim of this study, thus, defining the scope is in relation to these goals. This work examines

activity patterns, such as activity location and characteristics as well as travel characteristics; however studying route selection is beyond scope since the focus is on activity patterns rather than travel patterns. The aim of this study was to demonstrate that the travel and behavior patterns of a single target traveler could be learned quickly for prediction of that *same* target traveler's future behavior. While the testing methodology examines this across multiple travelers to ensure statistical significance; any analysis of the similarities or differences between the models of the individuals was outside the scope of this work. Likewise the models created of a traveler are traveler specific and any generalization of these patterns was not within scope.

It should be noted that all experiments within this study have focused on behavior within metropolitan areas. While the population of the different cities very significantly, from Anchorage with 362 thousand to the greater Chicago area with over 9.5 million people, all tests focused on urban areas (Bureau, 2008). Due to this, the results presented in this work may not be generalizable to rural areas where the types of activities available and distances to activities may be wildly different than those commonly found in urban areas.

Also, this work has focused on patterns within the United States and Canada. While further investigation would be necessary to confirm this, the techniques discussed in this work are not country specific and therefore may be able to be applied to other countries as well. It should be noted however there are known differences in travel behavior patterns even between developed countries with the same language due to differences in cultural attitudes towards things like the use of public transportation and variances in urban form (Giuliano and Narayan, 2003).

Therefore additional study would need to take place to explore the ability for travel patterns to be transferred across countries.

1.7 Definitions of key terms

Definition 1 ***Activity patterns*** *– The sequence of events and associated traveler context an individual makes with respect to the activities that occur and the associated travel to the locations where the activities take place.*

Definition 2 ***Traveler context*** *– A multi-faceted snapshot of characteristics of the activity and travel associated with the activity. This context is made up of a collection of information about the flexibility, planning horizon, and characteristics of the trip or activity itself. A sequence of this combined set of data is what defines the activity patterns. Urban planners have theorized a better understanding of how this context affects future behavior and what aspects can be shared across travelers is likely to not only produce better predictive models, but also identify patterns that can be transferred across users and locations (Timmermans, 2005).*

Definition 3 ***Target traveler*** *– Refers to the individual whose patterns are trying to be learned and predicted for. For example if the behavior of a number of travelers is being used to predict the behavior of the target traveler; this means that the behavior of other travelers is being used to predict behavior of the specific individual traveler that is the focus of the learning effort.*

Definition 4 ***Precision*** *– Mathematically described in Chapter 3, intuitively this metric captures the accuracy of a prediction when a prediction is made.*

Definition 5 ***Recall*** – *Mathematically described in Chapter 3, intuitively this metric captures the coverage of all items needing a prediction.*

Definition 6 ***Quality of prediction/predictability*** – *A term that loosely refers to how "good" the combination of precision and recall is for the prediction model.*

CHAPTER 2

LITERATURE REVIEW

This chapter presents an overview of work related to the topic of this study. A review of similar studies and contributions made by earlier works is presented to set the context of this research. As the approach used in this work is to use current models of traveler activity behavior and combine it with computational methods to allow modeling of an individual, this chapter is organized in that manner. It begins by discussing past transportation planning approaches to modeling activity patterns, followed by computational methods relevant to this learning task. The discussion of relevant works is organized into three main sections: traveler modeling, data collection methods and learning approaches.

2.1 Traveler behavior modeling

The focus of this section is reviewing different ways studies have approached modeling travelers. As there have been a number of different goals for various studies, it is not surprising that there are also many different approaches to this task. Here these various approaches are analyzed by the different ways travelers have been modeled in past studies and how these relate to the approach introduced in this work. The analysis of these studies is broken into three main sections: methods that have been used to model the behavior of travelers, techniques for generating activity patterns based on scheduling models, and methods used to model individual travel prediction.

Over the years, the type of data that has been collected for modeling travel behavior has changed dramatically. The first modern model of trip forecasting was introduced by Mitchell and Rapkin in the 1950s, which integrated travel and land use data to predict the amount of travel between areas within a city (Mitchell and Rapkin, 1954). This work became the basis of the four-step model that has dominated urban planning. The essence of this approach is to look at the number of trips between regions within a metropolitan area known as transportation analysis zones (TAZ). Data gathered for this approach has traditionally been paper-based travel surveys that capture the number of trips to the various regions in the area and when these trips occurred. While this has proven to be a useful tool in forecasting trips at a highly aggregate level, its level of predictions and lack of insight into the reasons for travel make it unsuitable for modeling the details of why travel was made. This realization led to a shift from modeling travel to trying to model the behavior of travelers to gain this additional insight.

2.1.1 Activity modeling

In the 1970s researchers began looking for ways to understand the reasons for travel referred to as activity modeling (Hensher and Stopher, 1979). Rather than just examining the trips made, a more behavioral approach was used to model trips based on the activity needs of travelers. Damm and Lerman introduced the theory behind how activity behavior could be used to determine the derived demand of activities for a person (Damm and Lerman, 1981). This work became the model of how activity demand could be used to determine activity scheduling, and thus the order of activity and travel could be based on the theoretical demand (Kitamura, 1988).

The model of the traveler was further extended by Recker *et al.* recognizing that scheduling behavior was not based solely on activity demand, but rather required a more complex model that accounted for the constraints of the various activities and the associated planning process (Recker et al., 1986). This belief that the planning process is tightly coupled with the actual scheduling of activities has led to much further study on how different types of activities are planned and adjustments to these plans. To build this type of model requires collection of not only activity characteristics, but also the activity constraints in terms of accessibility, location, and timing in order to create a planning model. While this type of model has significantly more data requirements, the result is more realistic model of when activities are likely to occur for simulating travelers in determining transportation demands.

In the early 1990s, Axhausen and Grling introduced a conceptual frame work for integrating the various components of activity modeling for the purpose of travel analysis (Axhausen and Grling, 1992). By the mid 1990's, McNally showed how these activity-based models could be used in micro-simulations to generate likely activity and travel patterns by micro-simulating individuals for the purpose of generating travel demand forecasts (McNally, 1996). Wen and Koppelman also introduced a framework for generating activity travel patterns by first generating stops for the traveler and then creating tours based on these stops (Wen and Koppelman, 2000). However, it was not until Bowman and Ben-Akiva's work in 2001 that a way was identified for these models to be used for activity scheduling and its implications for large scale travel demand (Kitamura, 1988; Bowman and Ben-Akiva, 2001; Vovsha et al., 2004). Since these

advances, the interest in this type of modeling has greatly increased as a potential alternative to the widely used four-step model.

2.1.2 Activity scheduling

Recent work in activity modeling has begun to focus on activity scheduling as a way to gain insight into understanding travel behavior (Jones et al., 1990; Axhausen and Grling, 1992; McNally, 2000). New data collection techniques such as the work by Doherty and Miller have focused on improving survey methods by capturing data on scheduling decisions and GPS traces to enhance researcher's insight into observed travel patterns (Doherty and Miller, 2000). This enhanced data can provide insight into modeling both how activities are scheduled, as well as how these plans change over time (Doherty et al., 2001; Doherty et al., 2002). The proposed research extends these ideas by attempting to discover activity scheduling patterns and adjustments to these schedules when events change over time.

Several studies have examined the scheduling patterns of individuals and how these might help in better understanding the patterns of travelers on a day-to-day basis. Bowman and Ben-Akiva's work has analyzed generating activities for individuals at a disaggregate or micro level, however the fit of these patterns to that actually observed in survey data was evaluated at a macro level (Bowman and Ben-Akiva, 2001). Susilo and Kitamura work analyzed the 6 week Mobidrive study showed that the centroid of activities and activity locations of individuals on weekdays is fairly stable, but variance is more likely to occur on weekends (Susilo and Kitamura, 2005). Other work has examined how shifts in the timing of routine activities can impact the scheduling of other activities resulting in a ripple effect of disruption and associated

rescheduling behavior (Kitamura et al., 2006). These findings are promising for providing meaningful predictions when a person deviates from their routine. An assertion of this work is that while it may sometimes be difficult to predict the exact next location, being able to predict the underlying activity is more insightful when there are multiple locations an activity may take place. While a location may not be constant, the proximity of the location of an activity pattern is relatively stable meaning different attributes of the predicted next activity are also likely to be useful. These works also address the importance of reevaluating the scheduling model continually to adapt to disruptions to the individual's routine and likely adjustments they will need to make in their final schedule.

This work extends these ideas by building upon the planning behavioral model. Since the purpose of this work is to learn the actual behavior of an individual, being able to generate a realistic activity order is essential. While previous studies have focused on how this type of model can generate realistic schedules for a simulated traveler, this study instead demonstrates how this same approach can be used to model the activity patterns of a specific individual.

2.1.2.1 Utility driven scheduling

Another method used to generate scheduling models is using a utility based approach. The basic form of these models is to determine the utility of different activities and scheduling the activities in the order of highest utility first. The primary differentiator for approaches in this area are the way utility is calculated for the activities and travel being modeled, and adjustments in the timing of activities to accommodate other activities.

In many works there is a separation of the travel utility and the activity utility with the basic assumption that the activity with the highest utility will be scheduled next and the choice of locations and thus travel will be made independent of the activity selected. In modeling the location choice and travel selection, several methods have been used. Ashiru *et al.* asserted that the utility of activity participation was driven by the utility of the activity combined with the individual's perceived cost of associated travel using a space-time route benefit measure (Ashiru et al., 2003b; Ashiru et al., 2003a).

Another approach to generating utility based schedules has been to generate schedules that optimize the amount of utility for a given day (Bowman and Ben-Akiva, 2001; Charypar and Nagel, 2005; Flotterod and Nagel, 2006; Janssens et al., 2006; Joh et al., 2001; Meister et al., 2005). This approach evaluates different ways of scheduling activities and travel to come up with the optimal utility for the complete day, which may not necessarily include the activities with the highest individual utility. One of the criticisms of this type of model is that it is unlikely the average person will think of the maximally optimal schedule, its easy to imagine this would be particularly true if doing so would change their regular routine (Bates et al., 2001).

Another approach that has been used is to consider not just the activity and travel utility but also account for the utility of the duration of the activity (Ashiru et al., 2003c; Ashiru et al., 2004). The main idea of this approach is to consider the utility of an activity not as a constant in scheduling, but to instead model the utility of an activity degrading as the duration of the activity extends. This model is then used to schedule activities and associated travel

in a utility order first approach. Additional activities may then be scheduled between these activities potentially shortening the duration of the previously scheduled activity (Joh et al., 2005). The main benefit of this approach is that the scheduling of activities is driven in a dynamic way that accounts for individuals trying to make the best use of their time based on all the activities they would like to accomplish in a more fluid fashion that likely better reflects real behavior. Ettema *et al.* extended this basic idea to also determine the timing and duration based on utility preferences for time of day (Ettema et al., 2007). Their approach was to use bell curves of time of preference to increase the utility of an activity or decrease the utility of an activity based on its proximity to the preferred time of an activity. Other work has combined these factors by considering the utility of scheduling an activity and associated travel based on activity need and value of time (Arentze and Timmermans, 2007).

While these models are well suited for generating schedules of simulated individuals that can reflect observed distributions, they also require a detailed model of the utility of the traveler. With simulated travelers this model of the utilities can be calculated by fitting the values to that observed in a large scale survey, however this becomes more complex when the target is an individual. If the target of the model is the schedule of a specific individual, this would mean determining the utility values specific to that individual. Since the utilities of activities vary to some extent by person, this would mean that a significant amount of observations of the individual would be required just to determine how often particular activities should be scheduled for that person. When considering a goal of this work is to determine the order of the activities on any given day this would mean developing a far greater knowledge of the traveler's

schedule and accessibility constraints and thus more data collection time. As a result, this type of model is impractical for learning the utility model of the individual quickly. However, the approach of using scheduling utility is still used but in a less complex form. Specifically one of the key approaches in this work is to incorporate the scheduling preference in terms of the order of activities. While considerably less complex than a mathematical model of the travelers utility, it does provide a way of quantifying scheduling preference with greatly reduced data requirements

2.1.2.2 Behavioral process driven scheduling

Behavioral process scheduling takes a similar form to that of utility based scheduling, however it attempts to model the schedule of activities not only based on the need for an activity but also based on observed scheduling behavior (Timmermans, 2005; Joh et al., 2005; Timmermans, 2005). The crux of this type of model is to simulate the way activities are scheduled and rescheduled as new activities are added. The resulting difference in this type of model compared to utility models is that the timing and duration of activities is fluid as additional activities are incorporated into the person's schedule. This difference in approaches to scheduling comes from modeling people's observed scheduling behavior as an ongoing task rather as a fixed schedule once generated. This idea is an integral part of adapting to unexpected activity choices made by the traveler.

The theory behind behavioral scheduling modeling began around 1980, and was later incorporated into simulations in Ettema *et al.*, which first established being able to use models of the scheduling decision process to simulate the activity schedules of hypothetical travelers (Hen-

sher and Stopher, 1979; Damm and Lerman, 1981; Ettema et al., 1993). The ideas behind this theory were later combined into a unified modeling framework in Doherty, and Doherty and Axhausen (Doherty, 1998; Doherty and Axhausen, 1999). This framework establish three main phases to the activity scheduling process: agenda formation, routine scheduling and the weekly decision process. This combined with more detailed data collection of the planning and decision process led to various forms of this conceptual framework forming the basis for many of the studies in this area (Doherty and Miller, 2000; Doherty and Miller, 2000; Doherty et al., 2002). Other approaches such as the work by Arentze and Timmermans recognize that travelers do not have perfect information and address how mental maps and cognitive learning affect the decisions made regarding activity and travel choice in micro-simulations. (Arentze and Timmermans, 2005). Joh *et al.* demonstrated how a utility based approach could be combined with the scheduling and rescheduling behavior to reorganize generated schedules to result in a higher utility than utility schedules that just added based on maximal utility of remaining activities by adjusting order and duration (Joh et al., 2005).

Due to the order items were considered in scheduling and rescheduling behavior potentially effecting the resulting end schedule, models of how planning decisions were made quickly gained attention. Some models addressed this by sequentially selecting an activity then determining the location (Doherty and Miller, 2000; Doherty et al., 2002). While work used a combination of the activity need combined with activity characteristics to determine the planning order (Kitamura et al., 1997). These ideas were extended with the SCHEDULER framework resulting in an ordering where high priority activities were added first and then an attempt was made to add

activities with less priority around the higher priority items (Grling et al., 1994; Kwan and Golledge, 1995; Grling et al., 1998).

In the area of micro-simulation, related work has primarily focused on using activity survey data for generating simulated activity schedules or verifying simulations (Pribyl and Goulias, 2005; Lee and McNally, 2003). Two micro simulators built based on the basic concepts of activity need, scheduling and rescheduling decisions are the Albatross simulator created by Arentze and Timmermans and the Toronto Area Scheduling Model for Household Agents (TASHA) by Miller and Roorda (Arentze and Timmermans, 2000; Miller and Roorda, 2003). The scheduling approach Albatross used was based on all mandatory activities being added before discretionary activities. The TASHA simulator by contrast took conventional trip diary data and added the activities sequentially using a fixed activity type order. TASHA would then build 24 hour activity schedules for micro simulation. The TASHA simulator demonstrated how a scheduling based model could generate realistic schedules at an aggregate level for trip rates and activity chain characteristics.

More recently researchers have determined that scheduling decisions are not as simple as a fixed planning order. One of the paths recent research in this area has focused on is shifting from predicting schedule sequencing in general to the scheduling and rescheduling decisions made in the planning process and its effect on the ultimate activity schedule (Joh et al., 2001b; Frusti et al., 2003). Pendyala and Bhat examined the relationship between timing and duration of activities and found that for non-commuters, activity duration was causal in determining activity timing (Pendyala and Bhat, 2004). Studies by Lee and McNally showed that activities with the

longest duration were scheduled first and then shorter activities that could be incorporated in a trip chain had the next priority (Lee and McNally, 2003; Lee and McNally, 2006). Doherty and Mohammadian examined the planning time horizon of activities and noted that even activities of the same type did not always have the same planning horizon (Doherty and Mohammadian, 2003). To address this their work examined using an artificial neural network (ANN) to model the planning time horizon actually observed in their study. Later work by Mohammadian and Doherty extended these findings to account for the high interdependency between activity and travel choice in scheduling and noted that situational factors can play a significant role in these decisions (Mohammadian and Doherty, 2005; Mohammadian and Doherty, 2006). Arentze and Timmermans introduced a different approach to the problem by starting with a base activity need value as in the fixed value studies, but then incorporated a time based component. Their approach adjusted the need value based on how long a task had been put off. Thus, the longer the activity went without being scheduled the higher the priority as the need became more pressing and its impact on household scheduling in addition to individual scheduling (Arentze and Timmermans, 2007; Arentze and Timmermans, 2009).

An area of study directly related to the study of planning order is how rescheduling occurs. Ruiz and Timmermans found that the impact on the existing pre-planned activity schedule and rescheduling to accommodate new activities was largely dependent on the types of activities involved (Ruiz and Timmermans, 2006). Nijland *et al.* found that the most common adjustment in rescheduling was shortening activity duration and depending on the relative reduction in time the activity may be canceled (Nijland et al., 2009). Auld *et al.* noted that how activities and

their timing was rescheduled was not dependent on the activity type alone (Auld et al., 2008). They identified the types of rescheduling that may occur when a new activity is scheduled including moving an activity forward, back, insertion or deletion.

While there has been extensive investigation in this area, it is important to note that none of these studies verify the scheduling or activity chains generated against real individual travelers. While several studies have examined the generated activity orders created in micro simulations, the accuracy of these schedules has been verified at an aggregate count level and that common trip chains are represented in the output. These methods verify the overall travel demand, however they are insufficient in verifying the schedules of real travelers. One of the reasons this occurs is that travel and activity routines vary greatly from person to person. Without accounting for these individual routines and personal scheduling dependencies a model is not going to be able to capture the actual behavior of an individual. A primary contribution of this work is extending these models to not only account for scheduling and rescheduling behavior but also account for personal routine in the form of sequential preference. While previous work has examined sequential ordering before, these studies have examined this only in terms of sequential dependencies rather than sequential preference (Janssens et al., 2005). As shown in Chapter 4, by accounting for individual sequencing preference, schedules can be generated that not only reflect aggregate counts, but also closer represent the patterns seen at an individual level. This finding is not only significant within the context of predicting the activity patterns of an individual, it also represents an advance in better simulating real patterns in transportation demand micro simulators.

2.1.3 Generation methods and evaluation

The value of using association rules to model the number of closely related attributes to activity characteristics and their relationship to activity scheduling behavior has been demonstrated in past work. Arentze and Timmermans work on creating the Albatross transportation simulation system showed that a model using a rule based system was highly effective at describing the complex relationship of mode, when, where and the duration of activity patterns (Arentze and Timmermans, 2000). Keuleers *et al.* extended these findings by showing that traditional choice modeling techniques compared to association rules were insufficient at modeling the combination of different attributes while still maintaining the spatial temporal patterns seen in multi-day activity surveys (Keuleers et al., 2001). Other studies have compared various learning methods to decision trees based on these types of rules, and while their aggregate sequential prediction shows promise, none of these studies have examined the quality of predictions at a micro level (Arentze et al., 2001; Janssens et al., 2004; Janssens et al., 2005; Joh et al., 2001). This distinction is important as the idiosyncrasies of the sequence of activity characteristic of an individual often do not follow the probabilities seen across the aggregate.

Evaluating the quality of generated activity patterns is a significant challenge due to the high correlation both intra set and across time. Many different measures have been introduced to evaluate this problem of multi-dimensional sequence alignment, however all of these techniques evaluate the quality of the generated sets based on a weighting established based on their similarity to commonly observed patterns rather than a specific instance (Joh et al., 2001a; Joh

et al., 2002; Shoval and Raveh, 2004; Shoval and Isaacson, 2007). The effect of this basis on commonality is less frequently observed patterns that deviate from that commonly observed are not accounted for and result in a biased measure favoring the more commonly seen alignment. The result is a suitable measure for evaluating the quality of a pattern generated compared to those commonly seen across a population, however evaluating the idiosyncrasies of an individuals behavior are lost. To account for this difference in goals this work uses the metrics of precision, recall, and F-measure commonly used in information retrieval studies. These metrics instead focus on the match of the predictions compared with those that actually occur on an individual basis rather than weighted based on those commonly observed, which we contend better represents the goals of prediction of an individual. This difference is explored further in Chapter 4.

2.2 Data collection methods

To better explain observed behavior, Stopher introduced the idea of activity diaries to replace travel diaries (Stopher, 1992). Prior to this work, while activities were being captured the way data was collected was primarily travel based. The fundamental difference of activity based diaries was to shift the focus of data collection to the activities and their characteristics rather than the travel as the primary factor. This shift captured a much greater level of detail of why the location they traveled to was chosen rather than just travel associated with getting to an activity location. From these diaries models were developed that focused on the reasons for travel and how these reasons influenced travel scheduling behavior.

Over the years the methods for collecting activity-based diaries and the type of data collected have made several advances that aid in the ease of data collection and ways of better understanding the planning process of travelers. Computer-based methods for collecting activity diary information have become much more common and in various forms (Doherty, 1998; Doherty and Miller, 2000; Doherty and Miller, 2000; Lee and McNally, 2001; Lee and McNally, 2003). This change has allowed researchers to get information on a participant's trip reporting on a more timely basis. Another advantage of this approach has been that it makes it easier for participants to enter their projected schedule and make changes to it allowing researchers to observe how people's activity schedule changes over time (Lee and McNally, 2003). Doherty and Miller demonstrated this type of approach could be used to record how participants activity schedules changed over a one week period (Doherty and Miller, 2000; Doherty and Miller, 2000). This study marked a significant advance as it recorded detailed information of the participant's activity plans and they evolved over time. By recording these adjustments at a detailed level over a one week period, a much more detailed view of rescheduling behavior over an extended period of time was able to be recorded.

Some of the more recent trends have been to augment the participant's manual entry of activity and travel information with GPS information. The first studies of this kind collected GPS information from participants in addition to travel diaries to validate the manual entry and provide additional data on route information (Murakami and Wagner, 1999). Using GPS logs also has had the advantage of being able to collect detailed information of where travelers went and how they arrived there with little additional burden on the survey participants (Stopher

et al., 2008a). This has allowed surveyors to analyze the GPS logs to help identify activity stops that were not being reported manually, thus improving the quality of the data (Doherty et al., 2001; Stopher et al., 2005a). Going a step further, Wolf *et al.* examined the possibility of using GPS logs to completely replace activity diaries (Wolf, 2000; Wolf et al., 2001). Their work recorded the GPS traces of participant's cars over a three day period and compared paper based surveys to using GPS data alone and prompting participants for activity details. Their work showed that while using GPS trace data alone was not sufficient for getting activity information, the combination of GPS traces and prompted recall surveying was very effective at recording detailed informations about travel while greatly reducing participant burden. Other work by Bricka *et al.* found that GPS data was helpful in reducing the common problem of non-response compared to non GPS enabled surveys (Bricka et al., 2009; Richardson et al., 1996).

Recent work in activity based surveys has begun to use GPS data not just as a contributing part of the data being collected, but also have the data make the survey process itself easier. One of the ways this is done is through processing the GPS trace during the survey period and using the information gleaned from the GPS trace to form a skeleton of the participant's activity schedule. By already having a skeleton in place, the burden on the participant for completing the information necessary becomes just a matter of recalling information about the activity skeleton which is referred to as a prompted recall survey. Other work by Clark and Doherty examined using GPS data to automatically track activity rescheduling decisions (Clark and Doherty, 2008). The approach in their work was to collect the plans of participants before

the day began and then compare a passively collected GPS trace with the trips planned as a basis of asking questions about how the rescheduling occurred.

This work builds upon these approaches by introducing additional ways GPS data can be processed on the fly thus allowing applications like an on-line prompted recall survey enabled by being able to automatically process GPS traces. This same approach is taken a step further by introducing ways GPS processing can be used on the fly to identify significant locations, revisited locations, and resumption of travel

2.3 Learning individual travel and activity patterns

Learning the travel and activity patterns of individuals has received a considerable amount of attention from a variety of fields. This attention has been in part due to the variety of different applications that use aspects of this work. These applications range from maintaining cell phone connectivity to projecting travel demand to location aware applications. While there have been numerous approaches, they have generally taken one of these forms: spatial temporal based projection, location-based context, or some mix of these.

2.3.1 Spatial temporal based projection

The basis of spatial temporal based projection is given a model of a particular space, predict where in that space an object will be at a future point in time. Applied to travelers this means given a point in time predict the location that person will be in the future and potentially how they arrived there thus the inferred travel.

Several researchers have explored techniques for tracking and predicting traveler's movement patterns. A significant portion of this work has been related to tracking and predicting the

movements of cellular users. Liu and Maguire examined mining mobility patterns and movement history of individuals for short-term prediction (Liu and Maguire, 1995b; Liu and Maguire, 1995a). Their approach involved identifying patterns of regular movement and utilizing this information to create a Markov model for predicting a user's next move in a simulated grid space. Bhattacharya and Das examined developing cellular user's mobility profiles to facilitate more efficient tracking of real users with uncertainty (Bhattacharya and Das, 1999). Their work explored how recent movement history could be used for short-term location prediction. Thus, while these techniques were useful for projecting immediate travel, they did not address projecting the length of stops or projected movement that is separate than the movement that is currently taking place.

Other work has examined the predictability of a traveler's movement by estimating the entropy of a traveler. Several studies have examined determining the entropy of user's movements between cell phone grids and used techniques such as Expectation Maximization to determine what grid a person is likely to be in and travel to next (Eagle and Pentland, 2006; Song et al., 2010). These approaches have shown that at a high level traveler's routines are fairly predictable. Other work by Eagle and Pentland has examined decomposing travel behavior into principle components referred to as eigenbehaviors and used similarities in eigenbehaviors to project future travel (Eagle and Pentland, 2009). This approach is significant as not only does it help predict routine travel, but also recognizes common deviations to the routine and their impact on the remaining travel within the day. However, all of these studies analyze this predictability at a cell phone grid level with an average grid size of 3 square kilometers for the work

by Song *et al.* (Song et al., 2010). Thus while these studies show that patterns at a high level are fairly predictable, the level of granularity is insufficient for determining the actual locations visited not to mention the possible different activities that took place in that relatively large area. As a result, while these models are useful at predicting high level patterns such as home to work trips and vice versa, they are insufficient for capturing tours that take place within one of these regions.

Transportation planning has taken a similar approach but rather than the goal being to project immediate location based on movement; the goal has instead been to project travel over time to get to locations. Wang and Cheng introduced a model for examining activity-based travel as a spatio temporal dataset for querying and modeling transport demand (Wang and Cheng, 2001). Schönfelder *et al.* examined using long term car GPS traces to identify areas where activities occur over periods of the day (Schönfelder et al., 2006). This work looked at car GPS traces to build a model of where stops were made throughout the day and the spatial proximity of these stops day to day to build a destination choice model to predict when trips would be made and where the stops were likely to occur at a more generalized level for predicting transport demand.

2.3.2 Location-based projection

Another area of research has been identifying the activity and movements of a person based on their location or history of locations. The premise behind these approaches is that the location in space has a direct relationship to either where they go next or what they are doing. These approaches take a couple of different forms, where the main factors are location knowledge

and movement between locations. Some of the most common forms include location projection based on location history.

Abowd *et al.* examined this in the form of the application having knowledge of the location and trying to determine the interests of the user based on the current location and the past locations (Abowd et al., 1997). With this type of approach the locations of interest are known ahead of time and as a result can be descriptively notated. While this type of approach may work well for applications such as a tourist application where the locations of interest are known ahead of time; for most applications this type of a priori knowledge is not practical for everyplace a traveler might go.

On the other end of the spectrum studies have examined traces for the purpose of significant location identification and predictions of next location based entirely on an individual's prior GPS traces. Ashbrook and Starner introduced a technique for automatically determining significant locations from GPS traces based on a combination of clustering and applying distance based thresholds (Ashbrook and Starner, 2002). Related work demonstrated that a second-order Markov model built on these locations and the frequency of trips between these identified locations could produce significantly better than random predictions of the next location a person would visit (Ashbrook and Starner, 2002; Ashbrook and Starner, 2003). Their work went on to show that GPS traces of multiple users in the same region could be used to jointly identify and label significant locations. While these works represent a significant step forward; as the authors acknowledge, the model is not suitable for abrupt changes in routine such as changes in class schedules, as it would take an extended period of time for the model

to adapt based purely on frequency counts. The weakness of this type of completely passive approach is that without more detailed information, the reason for the movement or even the activity information are not available.

Another approach to this problem has been to learn the activity at a location to infer future activities at the location. Several studies have examined using activity history at a location to infer future activity at the same location. Marka *et al.* examined identifying the activity of an individual based on the person's GPS trace (Marca et al., 2002). The study examined how a GPS trace could be used to infer activity information based on previous location visits given some activity log history.

Liao *et al.* addressed this problem by building up a history of travel and labeling the activities so that the task could be treated as a supervised learning problem (Liao et al., 2004; Liao, 2006). In their work, a model was constructed using a GPS trace where stops of a time threshold were given an activity label by the participant. From these examples a hierarchical Markov model to predict next location and activity information was built. This basic approach was also explored for indoor patterns by using features such as availability of WiFi networks to determine activity locations. This technique was shown to be effective at predicting next destination and high level activity information for travel within the training set examples.

Other work moved from a supervised approach to an unsupervised approach to build a model based on a unlabeled GPS traces. This technique used expectation maximization (EM) to learn transition probabilities and identify significant locations such as goal destinations (start/end

points) and mode change locations (bus stops and parking lots) (Liao et al., 2007). The resulting model was capable of adapting end goal probabilities based on en route location information, and identifying significant deviations from expected travel as long as the trips were within previously visited routes and locations. So while more advanced than a model built strictly on significant location identification, it still would required some external entity to determine the activity or why a trip was made. In similar work, Gogate *et al.* introduced a hybrid dynamic mixed network approach to modeling individual travel that identified significant locations and projected next destination and route selection (Gogate et al., 2005).

While these works mark significant advances in moving from focusing on location alone, the approaches are still completely dependent on known locations. The significance of this is that deviations from these known locations result in the models not being able to provide any information about the current situation or context. This work differs from these approaches in that like the GPS based techniques, it focuses on individual behavior, but the goal is more similar to micro-simulation activity models, in that the interest of this work is modeling a more feature rich set describing the reason for the behavior beyond the locations. The result from this shift in approach is being able to still make reasonable predictions regarding the type of activity, likely duration, and mode of transit for the next trip even when a location has not been previously visited. The other benefit of this more behavioral based approach is being able to predict the activity needs of the person and travel constraints such that an intelligent mobile application could then suggest other places with that same activity that are still accessible to that person.

2.4 Learning methods

Within this work we examine using a combination of the activity based approach combined with the detailed level of examining individual patterns used in the location-based approaches. While this work looks at combining the ideas from both of these approaches, one of the goals of this work is to prevent the model from being tightly tied to previously visited locations which is a critical component of past location-based methods. To address this issue the abstraction of activity models is used at an individual level. As discussed above, one of the features of the activity based approach is creating a model of a set of attributes that represents the activity as well as the travel characteristics rather than just the destination. As a result, the model of the traveler is a sequence of sets of attribute values that are closely related both within the set itself as well as sequentially. This formulation of the problem was one of the driving factors behind the learning methods considered.

2.4.1 Associative and sequential mining

One of the key aspects that must be considered when modeling the sequence of sets that forms the activity sequence is the strength of relationships both within each set as well as the dependencies across the sequence itself. Sequential associative mining is well suited to this problem as it specifically addresses strong relationships both within the set as well as across sets. Since associative and sequential associative mining were first introduced by Agrawal *et al.*, this approach has been the focus of many studies (Agrawal et al., 1993; Agrawal and Srikant, 1995). Zaki introduced length, width, and maximum gap constraints to reduce data mining time. In Harms *et al.* and Harms and Deogun introduced an algorithm for mining frequent

episodes from multiple sequences using time lag constraints and separating antecedent and consequent constraints (Harms et al., 2002; Harms and Deogun, 2004). All of these works have focused on extracting and applying the associative rules uniformly across sequence series in prediction. In addition, all of these works tend to focus on predicting a single next item in a given sequence. Other work has examined set prediction as constrained sequence mining, but focused on traditional sequential association rules (Garofalakis et al., 1999). More recent work by Lahiri and Berger-Wolf has addressed periodic patterns and partially periodic patterns in constrained dynamic graph problems (Lahiri and Berger-Wolf, 2009). Their work on mining graph patterns in dynamic networks presents a promising way of evolving patterns over time, but as the missing value problem is a central issue in this work; the algorithm's basis on connectivity being either present or absent is not well suited for missing information. Liu *et al.* introduced a template based technique for rule mining referred to as label sequential rules. Applied to text mining, these rules were effective at being able to discern the context of a word and thus the applicability of a rule based on a comparison to a mined set of templates (Liu et al., 2005; Jindal and Liu, 2006).

Identifying patterns in traditional associative mining relies on training sets composed of multiple instances of sets for its primary constraint support. With associative sequence mining, there is a similar dependency on multiple sequences (Hipp et al., 2000). The implication of this is when applied to the context of transportation for a travel or activity pattern to be significant, the pattern must be present across multiple travelers. While this constraint is likely a good guideline for predicting traveler patterns in general, if the goal is to predict the travel

pattern of an individual, then patterns that are unique to that individual are likely significant for predicting future behavior of that individual even if they have little predictive value for the set of all travelers as a whole. In addition, these techniques are not well suited to lengthy sequences, as the distance between sets within a sequence is not accounted for. Applying this logic within the context of transportation, this would be equivalent to saying the likelihood of an event occurring is just as dependent on an activity that occurred four days ago as it is on the previous activity. While such relationships may exist, it seems reasonable to assert that activities that occurred in the traveler's recent history are in general more likely to be better predictors of the immediate next activity.

2.4.2 Missing values

Another aspect that had to be considered in selecting a learning method was tied to one of the common problem with activity surveys, the issue of non-response (Richardson et al., 1996). Besides non-participation, this issue takes two forms: not reporting an activity; or not fully answering questions about their participation in an activity. In traditional activity surveys, this poses a problem for under reporting activities and incomplete information. However, by looking at the aggregate of many histories, respondents that do provide this information can at least provide some coverage of the missing information. When learning the patterns of an individual, this becomes more problematic due to not having the advantage of coverage through the participation of others. These issues could cause significant problems for learning approaches since predictions would be made without knowledge of activities that took place and thus not recognizing the need to include this history in future predictions. The aspect

of missing information about an activity could mean a crucial piece of information for future predictions is not recorded. This is particularly a concern when selecting the prediction model to use. While some methods are effective with complete information, their predictive quality may degrade faster than other methods as the amount of missing data increases.

Handling missing data has been a challenge and a source of considerable research over the year's. As discussed in Rubin, one of the greatest challenges of this problem is that unless the assumption can be made that the occurrence and the values of the missing elements are truly random, missing at random (MAR), it is difficult to determine which factors are most important in determining the missing values (Rubin, 1976). To address this problem in general, one of the common ways this problem has been addressed has been iteratively filling in the missing values based on various ways of predicting the missing values given all the present values whether these were the original values or inferred values (Schafer and Graham, 2002; Raghunathan et al., 2001). One of the biggest challenges in this is that the identification of which are the most important factors can greatly affect the accuracy of values as an incorrect selection can continually cause later values to be populated poorly based on bad earlier selections. As Lakshminarayan *et al.* demonstrated in their work, since real data sets do not follow the MAR assumption, the algorithm used for one data set may not be the best algorithm for another data set (Lakshminarayan et al., 1996). While different techniques work better than others depending on the data set, some of the methods that have had good results in general for populating single records of data (i.e. data without temporal relations) are maximum likelihood (ML),

Bayesian multiple imputation (MI), and associative rules (Schafer and Graham, 2002; Ragel and Crmilleux, 1999).

With associative mining the aspect of incomplete information is of particular concern since its effectiveness degrades rapidly using traditional associative mining when there is incomplete or missing data. Addressing this problem has been a topic that has gained much attention. Ragel and Crmilleux's approach to this problem, known as missing value completion (MVC), was to create multiple views of the data so that when mining any attribute only the records that contained the attribute were included (Ragel and Crmilleux, 1998; Ragel and Crmilleux, 1999). The result was relationships could be effectively mined by only considering the relationships when a value did appear. One of the problems of this approach is that while it is very effective for associative set mining it does not work well for sequential associative mining. If this approach was applied to sequences, the only sequences that would be considered would be ones in which the attribute value of interest appeared in all sets of the sequence greatly reducing the likelihood of being able to identify more subtle patterns. Shen *et al.* examined a different approach to this problem based on using association rules and combining these with sub-frequent item sets to populate missing values (Shen et al., 2007). More recent work by Bashir *et al.* takes an iterative approach using a combination of association rules and k-Nearest Neighbor (kNN) to impute the missing values and improve the performance of the association rules in missing value scenarios (Bashir et al., 2009). The general approach of their method is alternate between filling missing values with association rules when possible and then when no rules are applicable, kNN is used to impute remaining values.

While many techniques have been introduced to handle the problem of missing data, most of these techniques are not suitable for sequential data. Within this work we extend upon work in this area by introducing an algorithm that is able to populate missing data effectively for both associative sets as well as sequences of sets. As shown in Chapter 4, the technique introduced builds upon the ideas introduced in these works and contributes a significant advance in being able to impute values that not only depend on its own record but also sets with sequential relationships.

2.5 Activity transferability

The transferability of activity models across cities has been a source of interest both in its theoretical implications as well as the practical benefits. Arentze *et al.* explored the theoretic transferability of activity patters across cities using the Albatross rule based model (Arentze et al., 2002). Their approach to evaluating activity transferability involved building a model on one city and using the model to compare the predicted travel demand from the resulting activity patterns. There work demonstrated at an aggregate level the predicted demand was satisfactory compared to models built specifically for these cities. While this result is significant, the ability to transfer these patterns in terms of scheduling patterns or validity against individual patterns was not explored. This work builds upon these findings by examining this same approach to transferability of activity patterns across cities however evaluating against actual observed behavior in the form of activity surveys rather than simulations. The second aspect that is an essential difference of this work compared to their work is examining the transferability in terms of the ability for activity patterns to be transferred across cities at a individual level.

This study thus expands on their work by also verifying the transferability of activity patterns at a micro level in addition to their findings at a macro level.

Other studies have examined the similarity in travel patterns across different countries (Timmermans et al., 2003). One of the interesting aspects of this study was how different cultures, urban structures, transportation networks and accessibility differences affect activity and travel patterns. This work found that differences in relative location and transport network were far less important than demographics, economics and social aspects. While this study demonstrated similarities in aggregate travel patterns across these cities, little analysis was made at a more micro level in terms of ordering of activities and interrelated constraints.

While these studies have shown similarities in patterns across cities, none of these studies demonstrate at a micro level whether the scheduling patterns of one city can be used to help infer individual scheduling patterns in a different city. This work extends these ideas by demonstrating that the micro patterns of one city can be used to help augment the observations within a different city for more accurate micro level predictions. While this work does not explore if this relationship upholds across significant cultural differences, it does show that these patterns can be transferred across varying transportation networks and urban structures within a selection of cities within the United States and Canada.

CHAPTER 3

METHODOLOGY

This chapter presents the research design and context of this work. Afterward, the data requirements and data sources used to address these needs are reviewed along with a description of the data collection effort conducted as part of this study. Next, the methods used to analyze the data are explained, followed by the procedures used to evaluate the findings. Finally, a summary of the methodology presented in this chapter is reviewed and an outline of the overall experimental approach is presented.

3.1 Research perspective and type

The goal of this work is determining how best to learn the activity patterns of an individual for predicting new or missing elements of their activity information. Since applications that are likely to use the generated predictions would primarily be interested in the measurable quality of the predictions, this study used a quantitative approach. While qualitative information such as whether the learning is effective or usable would also be beneficial, as the output of this study is likely an input into another application rather than an end goal, a quantitative approach is likely more meaningful. However, qualitative aspects are also used for softer measures such as model "understandability", learning "quickly", or what is a "reasonable" effort. For these types of concepts that are more difficult to quantify a more subjective approach was necessary, but even these will be compared objectively against existing methods.

Since we are primarily interested in a specific application, activity behavior, that is associated with a physical set of data, we examine the outcomes of this study using an experimental approach. This approach seems a natural fit given the goal of evaluating how well the algorithms and approaches introduced in this work will work on practical applications. Specifically the methods proposed in this work are evaluated against real data sets collected from travel survey participants to confirm if the projections hold true for real scenarios. To accomplish this goal, analysis was performed against data collected as part of this work as well as travel surveys collected in other studies. These data sets became the basis for both training the models as well as evaluating the resulting models. Thus, all outcomes of this study have been experimentally verified against "real world" data.

3.2 Research context

Historically analysis of activity-based travel patterns have been studied almost exclusively for modeling behavior of a person based on a set of demographic characteristics (Algers et al., 2005). These models have then been used to project the travel of an area by simulating the behavior of a population based on the demographic makeup of the group with the results verified at an aggregate level. Furthermore, those interested in the resulting patterns and projected behaviors have almost exclusively been urban planners and researchers, but this is ripe for change. This change is being driven by several closely related factors: better technology, more complete data, and additional interest.

The data available for transportation studies in the past has by in large been limited to aggregate traffic information and paper-based travel surveys. With the increase in prevalence

of hand-held computers and GPS devices, the detail and types of information that can be captured for travel histories are becoming far richer than ever before. While these data sources share many characteristics with other studied travel activity logs, the additional information on decision processes and opportunities for further enhancement with GIS offer some unique opportunities for behavior analysis. In addition, the granularity and quality of data being captured related to route/location choice and time in particular is far superior to data collected in the past. Many of these problems require new algorithms and techniques to fully realize the benefits of the new data sources. The capture of these extra details offers additional possibilities in identifying models of behavior that are not just sensitive to demographic data and stated preferences, but also to observed similarities in travel preferences.

The spread of hand-held computers and GPS devices are also playing a significant role in changing past trends of who is interested in transportation information. With the projected huge increase in market penetration of GPS-enabled PDAs and cell phones, the market is ripe for the emerging fields of location-aware and activity-aware applications (Research, 2009; iSuppli, 2007). The spread of these devices allows the movement of people to be observed and utilized in ways that were previously impractical. This phenomenon provides a new realm for applying travel behavior analysis in real-time rather than ex post facto as it has primarily been done in the past. Applications that can take advantage of this combination of factors for the device owner's benefit seem a natural extension as evidenced by the huge growth in location based applications. The proposed research intends to address this space by providing travel behavioral

analysis and projection of the individual for their own benefit that other applications can utilize as a service for applications such smart traveler assistants.

The primary goal of this work is to quickly learn the travel and activity patterns of individuals in a reasonable manner. To accomplish this task we examine three sub-goals as part of the proposed research:

- introduce techniques necessary to collect the necessary history of the individual in a relatively unobtrusive manner over the long term
- design algorithms to maximize the value of the data that is collected, and
- develop techniques for mining transferable travel behavior patterns for the purpose of speeding up the learning of patterns of the target individual

These parts will culminate in a framework introduced for combining these elements in an end-to-end solution.

3.3 Data requirements and collection methods

The purpose of this section is to explain the data needed by this study to validate its findings as well as explain how the sources used were selected. First, the data required for this study and its purpose is explained. Second, the rationale for selecting the specific data sets used in this work that are taken from other studies are explained as well as their limitations. Finally, the gaps between these existing data sources and that needed by this study are discussed, leading into the motivation for the additional collection effort that was made as part of this work. A description of the survey that was conducted is explained along with the goals of the survey.

3.3.1 Review of data requirements

To determine the information that needed to be collected and the method of collection for this study, there were a couple of factors that had to be considered. The overall purpose of this work was to learn to predict the activity patterns of an individual. Within this work activity patterns are defined as the location the activity takes place, the activity characteristics, the travel characteristics associated with getting to the activity location, and when these items are planned. These categories were selected as they capture information that could be used by applications such as an intelligent traveler's assistant or personalization related to the traveler.

Most existing studies of projecting the patterns of individuals have examined learning traveler behavior by concentrating on the travel itself and some have examined trying to determine the activities of the person based on location alone. Instead of examining the traveler by looking at where they go and how they get there, this study takes a different approach by instead looking at the decision factors and reasons behind the travel. Transportation planners have studied travel behavior extensively over the years. More recently, their focus has shifted from looking at travel alone to understanding why the trip is made and when this decision was made (Timmermans, 2005). One of the approaches used for this has been examining the activity needs/desires of the person as the reason the travel is made. Based on the observation that information about the activity and travel options drives the decision of where the activity will take place and how the traveler will choose to get there. We propose this type of information is more important than travel and location information alone in understanding what information is useful to the traveler *before* they make their trip.

To accomplish this, we chose a set of activity and planning data similar to that traditionally collected by transportation planning activity based surveys. The reason for this choice was in part due to the proven record of this type of data being effective at modeling traveler behavior. The validity of using this type of activity data to create multi-user tour based scheduling models has been demonstrated in recent work (Doherty and Mohammadian, 2007). A second reason this model was selected was due to the minimal changes necessary to adapt to learning an individual's behavior compared to the traditional use of modeling a type of individual. Finally, the fields chosen captured the key attributes of interest while still allowing several of them to be collected via passive means such as a GPS trace. By choosing to focus on many fields that could be collected via passive means to build a history of the user's behavior, it allows a model of the traveler to be created and refined with little burden on the traveler. Still other fields selected, while not addressed directly in this work, can either be drawn though integration with other applications such as an online calendar or have shown promise for being able to be collected passively in other research, such as mode of transit (Stopher et al., 2005a).

The data selected for use in these experiments was activity, location, mode of transportation, arrival time, departure time, duration, sequential order, and planning flexibility information. These 10 attributes listed below describe the traveler context examined in this work and all have discrete values:

- **Observed sequence of the activity** - integer
- **Location** - location id (explained later in Section 3.4.3)

- **Activity** - 22 values: 'At home other', 'Shopping grocery', 'Social', 'Religious/Civic', 'Leisure/Entertainment', 'Meal', 'Shopping household', 'Household errands', 'Recreation', 'Shopping other', 'Personal business', 'Pick-up/Drop-off', 'Shopping major', 'Health care', 'Work/Business', 'At home work', 'Primary work', 'Services', 'Other', 'Volunteer work', 'Change transportation', 'School'
- **Mode of transportation** - 10 values: 'Auto-drive', 'Walk', 'Auto-pass', 'Multi-modal', 'Other', 'Commercial rail', 'Bus', 'Bike', 'Light rail', 'Taxi'
- **Arrival time** - 8 values: '3am-8am', '8am-10am', '10am-12pm', '12pm-2pm', '2pm-4pm', '4pm-6pm', '6pm-8pm', '8pm-3am'
- **Departure time** - 8 values: '3am-8am', '8am-10am', '10am-12pm', '12pm-2pm', '2pm-4pm', '4pm-6pm', '6pm-8pm', '8pm-3am'
- **Duration** - 7 values: '10 minutes or less', '10-30 minutes', '30-60 minutes', '1-2 hours', '2-4 hours', '4-8 hours', 'Greater than 8 hours'
- **Duration flexibility** - 3 values: 'Inflexible', 'Somewhat flexible', 'Very flexible'
- **Time flexibility** - 3 values: 'Inflexible', 'Somewhat flexible', 'Very flexible'
- **Spatial flexibility** - 3 values: 'Inflexible', 'Somewhat flexible', 'Very flexible'

These attributes were selected as they represent a mix of information about the type of activity, location, relative time as well as actual time the activity took place, and the person's flexibility. By structuring the data in this way, a trace of the discrete events of the traveler can be thought of as a stream or sequence of events with the set of attribute values at each

event being highly related. To address the end goal, a number of data sources from both other studies as well as this study were used to demonstrate the findings of this work.

3.3.2 Experimental data selection

To accomplish the goals of this work there were a number of different types of data necessary to evaluate the algorithms introduced in this work and verify the subsequent findings. First, as this work aims to not only predict the information about the observed activity and travel behavior, but also the planning process; a data set that captured this combination of features was necessary. Specifically a data set was needed that captured detailed planning data in conjunction with activity and travel information. Second, as one of the goals of this work was demonstrating that activity and planning patterns from one city could be used to improve the learning time and pattern coverage for another city, this type of data was needed from at least two cities. Third, to demonstrate the validity of this work large activity surveys were needed from multiple cities. Finally, a data source that captured a GPS trace in conjunction with an activity-based survey was needed to demonstrate that passive data sources could augment and reduce the need for active collection mechanisms.

The data sources included in this study reflect an effort to maximize the limited data available specific to this study's internal goals, while still demonstrating its external validity. Specifically the combination of data required for all aspects of this work, to the best of the author's knowledge, can only be found in the survey conducted as part of this research. As a result, a number of sources outside of this study have been incorporated in various ways to demonstrate the generalizability of this work despite the limited amount of data available.

Described below are the data sets used from other studies to meet these requirements, followed by the survey executed as part of this work.

3.3.3 Existing data sources used

To fulfill the requirements stated above, additional data sources were needed from existing studies to supplement the data collected as part of this work. For evaluation, this thesis focused on three sources of outside data. The purpose and use of each of these data sets is described briefly here, but its exact usage varied depending on the goals of the particular experiment. The details of how a specific data set was used in an experiment is defined in Chapter 4. Described below are a summary of the pertinent information of each of these data sets, with references to the appendix for more detailed information.

First, a large activity dataset that contained detailed planning information is described. As the additional time required of participants for this type of survey is often burdensome, the size of these planning studies tend to be smaller than most standard activity-based surveys. This is one of the reasons additional surveys were necessary to confirm the validity of this work. To address this need for larger datasets and data from multiple cities, two additional surveys were also included, which are described afterwards.

3.3.3.1 Computerized Household Activity Scheduling Elicitor survey

To address the need for an activity-survey that also included planning data, the Computerized Household Activity Scheduling Elicitor (CHASE) survey was selected as it provides a large study of detailed planning data in conjunction with activity and travel information. A secondary reason for selecting this study was that the survey collected traveler histories over a

seven-day period, which allowed a more extended view than many surveys that only captured two days.

The CHASE survey was conducted in Toronto, whose metropolitan area contains 5.5 million people, between 2002 and 2003, it consists of travel activity information for 426 adults in 271 households (Bureau, 2001; Doherty and Miller, 2000). This survey focused on observed activity and travel patterns as well as the planning process associated with these activities captured over a seven-day period for each individual. The data was collected via a computerized scheduling program that allowed survey participants to self-report on their actual activities as well as their on-going scheduling process. The user was asked to add, update, and delete their travel activity plans as their week progressed. The final state of the scheduled activities was recorded as the observed activity sequence similar to the activity sequences that are recorded in traditional surveys based on activity diaries. Additional details on the survey methods, sample characteristics, and data quality analysis can be found in (Doherty and Miller, 2000; Doherty et al., 2004).

The CHASE survey captured a wide range of in-home and out-of-home activity types. These activities are captured and organized by a high-level classification as well as a more detailed activity type. The high level activities that are examined in this thesis are organized into ten classifications (Night sleep - other needs, Social, Meals, Work/School, Household Obligations, Drop-off/Pick-up, Shopping, Services, Active recreation, and Entertainment), which are further broken down into 52 specific activity types (Grocery shopping, Internet shopping ... exercise, active sports, etc.)

In addition to the activity type, detail of the involved persons was also captured in the survey. For each activity, subjects were asked to indicate other people that were directly involved with the activity from a pull-down list of names. The list of names was initially populated via interview and the relationship with that person was noted. As the survey progressed, participants could add new people to the list by specifying their name and relationship (son, daughter, spouse, relative, friend, etc.). Parents of young children were asked to specify which of their children were "under their care", to obtain further detail about this participation. This information was used to further differentiate the three main activity types from above into five categories used in the most recent versions of tour models (Vovsha et al., 2004), including:

1. Individual mandatory activities (i.e. not conducted with other household members, but may be with other non-household members);
2. Joint maintenance activities (i.e. conducted with other household members)
3. Joint discretionary activities (i.e. conducted with other household members)
4. Allocated maintenance activities (i.e. not conducted with other household members, but may be with other non-household members)
5. Individual discretionary activities (i.e. not conducted with other household members, but may be with other non-household members).

Almost all activities fell into this categorization, except a portion of individual mandatory activities conducted with other household members.

This dataset contains a mix of information about the type of activity, the location, relative time the activity took place (# tour of the day), and the persons involved. Thus, the dataset can be thought of as a stream or sequence of events with the set of attribute values at each event being highly related. For this dataset, there are 41,312 sets of traveler context, with the average adult's traveler context sequence being just over 92 sets in length. Additional information about the data set can be found in Appendix A.1.

While the number of households surveyed by the CHASE study is substantial compared to other studies that capture planning data; it is considerably smaller in the number of households than more traditional activity-based surveys collected elsewhere. As a result, the CHASE data set is limited in being able to generalize the activity results overall. In addition, GPS traces are not captured as part of this survey. Thus, additional data sources were needed to meet the data requirements.

3.3.3.2 Metropolitan Travel Survey Archive

This section describes additional surveys that were chosen to make it possible to show that the models introduced in this work are neither city specific nor city size specific and that the activity patterns are generalizable. Since this work is also intended to show that patterns can be found across different cities of different sizes, large activity-based surveys of two different cities were selected to fulfill this need. While these datasets do not offer the same level of planning detail, the two selected surveys offer much larger data sets and the opportunity to verify geographic transferability.

The Metropolitan Travel Survey Archive[1], a publicly available collection of 29 metropolitan travel surveys, was used as the source for this data. To capture a range of different city sizes, the Atlanta 2001-02 survey and the Anchorage 2002 activity-based surveys were chosen. The benefits of these two surveys in particular, in addition to the variation in sizes, was both of these studies also captured similar details to the CHASE data set and this study's own collection effort making them easier to integrate. The details of the data sets selected to fulfill these needs are described in detail below.

Atlanta

The 2001-02 Atlanta Household Travel Survey conducted in the Atlanta metropolitan area, which has a population of 5.6 million people, was chosen for several reasons. This dataset contains a large number of households and attributes about activity and travel details at each step and their ordering, making it well suited for sequential traveler context prediction. Second, the type of activity and travel data collected in this survey is very similar to that collected in the CHASE survey. Finally, this data set represents one of the larger publicly available data sets of this type, making the results of this study open to competitive comparisons by other work in this area. Its large size allows for some experiments that would be impractical on smaller data sets. For example, demonstrating one of the proposed applications, learning given different quantities of data missing to reduce survey participant burden (Marca et al., 2002; Auld et al., 2009). These experiments require a large data set in order to show a significant portion of the

[1] http://surveyarchive.org/

data could be removed (i.e. a number of survey questions reduced) while limiting the impact on prediction quality.

The 2001 Atlanta Household Travel Survey was conducted from April 2001 through April 2002 by NuStats on behalf of the Atlanta Regional Commission (ARC) (NuStats, 2003). The data consists of travel activity information for 21,323 persons from 8,069 households and 126,127 places visited during the 48-hour travel period. This survey focused on observed activity type, timing, and travel associated with each person's activity schedule captured over a two-day period. The survey captured a wide range of in-home and out-of-home activity types, which were broken down by a high-level classification. The survey captures 255 attributes relating to the travel, activity, and demographic characteristics of the individual for each activity sequence that was recorded. The data is structured such that each event corresponds to an activity in the person's schedule with the set of attributes corresponding to the characteristics of the activity and travel associated with getting to the activity. Additional information about the data set can be found in Appendix A.2.

Anchorage

Like the Atlanta survey, the 2002 Anchorage Household Travel Survey was chosen for several of the same reasons. First, while much smaller than the Atlanta survey, it still contains a much larger number survey participants than the CHASE dataset. Second, the type of activity and travel data collected in this survey is very similar to that collected in the CHASE and Atlanta surveys. Finally, this data set was chosen as it represents a vastly different city size compared to the other data sets studied. While the metropolitan areas of the other cities studied are

very large: Toronto - 5.1 million, Atlanta - 5.3 million, Chicago - 9.5 million; Anchorage by comparison is much smaller at 362 thousand, and thus represents a good test of how transferable travel patterns are across city sizes (Bureau, 2008; Canada, 2008).

The 2002 Municipality of Anchorage Household Travel Survey was conducted from February 2002 through May 2002 by NuStats on behalf of the Municipality of Anchorage (NuStats, 2002). The data consists of travel activity information for 1,293 households and 7,545 addresses visited during the 24-hour travel period. This survey focused on observed activity type, timing, and travel associated with each person's activity schedule captured over a one-day period. The survey captured a wide range of in-home and out-of-home activity types, which were broken down by a high-level classification. The survey captures 91 attributes relating to the travel, activity, and demographic characteristics of the individual for each activity sequence that was recorded. The data is structured such that each event corresponds to an activity in the person's schedule with the set of attributes corresponding to the characteristics of the activity and travel associated with getting to the activity similar to both the Atlanta and CHASE surveys.

3.3.4 Urban Travel Route and Activity Choice Survey

While the sources listed above address many of the data requirements, none of them also captures the participants GPS trace. As described briefly above, one of the key contributions of this work is being able to identify travel and activity locations automatically is based on passive information collection like GPS data. While other surveys have captured GPS information as part of the survey process, few have captured the combination of planning data and multi-modal traces (Stopher et al., 2007; Schönfelder et al., 2006; Bricka et al., 2009; Clark and Doherty,

2008). To address this gap and collect additional data to evaluate the benefits of the ideas created as part of this work, we augmented these data sources with our own data collection effort. The result was the Urban Travel Route and Activity Choice Survey (UTRACS), which recorded activity information, travel information and planning information in addition to GPS traces as part of an internet-based prompted recall survey. Described below is the approach used to address these questions and how they were incorporated in the UTRACS survey conducted by UIC in conjunction with the Illinois Department of Transportation (IDOT). This section describes the basic information of how the UTRACS survey was conducted by these groups to illustrate the validity of using this survey to verify the findings of this study. The details of this design and implementation of the collection procedure are described further in work published related to this study (Auld et al., 2009; Frignani et al., 2010a).

3.3.4.1 Measure benefit of passive data collection

One of the challenges of evaluating the potential benefits of passive data collection was the lack of sufficient data in previous studies. Specifically being able to evaluate the value of these passive data sources with respect to a multi-modal activity study. While numerous activity studies had taken place, at the time of this study the combination of a multi-modal study combined with integration of GPS data was rare. The aspect of multi-modal studies is important as a pedestrian may be continually moving at a activity location whereas this is not likely to be the case with a car based study. One of the goals of the UTRACS survey to address this gap.

While there are ways GPS traces can be used in relation to an activity survey, the primary use examined in this work was how well GPS data could be used to identify significant activity stops. The approach to verifying this with the UTRACS survey was to collect the activity diary of the survey participants allowing them to confirm expected significant locations while being able to add, delete, or move the expected significant locations if they were identified incorrectly. Using this methodology, the accuracy of being able to identify significant locations from GPS traces could then be determined. Thus by comparing the number of correct identifications versus the number of missed or incorrectly identified locations the merit of the GPS processing algorithm could be determined.

3.3.4.2 Measure reduction in traveler burden

The second aspect that needed to be addressed as part of this study was a measure of the time required by participants and the associated reduction in time that could be expected through methods introduced in this work. To address this, in addition to the data recorded as part of the data entry by survey participants, the time it took participants to answer various components of the information asked was also collected. this could to collect various forms of data so that an analysis could be made of the time savings related to the types of information that could be gained by the techniques introduced in this work.

3.3.4.3 Survey participants

The UTRACS survey was conducted in the Chicago Metropolitan Area from May 2009 through December 2009. A total of 100 households in the Chicago area were surveyed. Respondents were recruited from a random stratified sample of the Chicago area population. Half of

the sample was constituted by individuals age 65 and over and the other half of ages 16 to 64. The geographical area included Cook, DuPage, Lake and Will counties. This sample was stratified by county and by four categories of income. The sample followed the geographic population distribution existing in Census 2000 (Bureau, 2001). However, because of past experience of lower response rate among lower income and lower education households, those falling in the lower income categories were oversampled to yield a final income distribution similar to that of Census 2000 (Mohammadian et al., 2009; Bureau, 2001).

Familiarity with computers was not required from respondents since the survey requires little computer knowledge, and training and assistance were thoroughly provided. For those households that did not possess a working computer and internet connection, laptops with dial-up or wireless broadband were provided. Assistants either left the laptops and the internet data card in the households for use during the survey period or they visited households taking the equipment every two or three days.

3.3.4.4 Instruments used to collect data

The equipment used in this survey consisted of GPS trackers, rechargeable AAA batteries, chargers and computers with internet connection. The GPS tracker had a storage capacity of approximately 360 hours of tracking data and could operate for 15 continuous hours. It weighed approximately 50 grams, not including batteries. Approximately the size of a cell phone, the device was selected as it could easily be carried by the participant throughout the day regardless of the mode of transportation. The GPS tracker recorded latitude, longitude, speed, and satellite information every five seconds. Code was placed on the device to automatically

process the raw data upon connection with the participant's computer. Once connected the survey participant would upload the processed log file to the survey web server for further analysis.

3.3.4.5 Survey execution

The incentives for participation were a 25 dollars debit card for each respondent in the household, and the entry into a drawing to win one grand prize of 500 dollars or one of two first prizes of 250 dollars, also in form of debit cards. Respondents were entitled to the 25 dollars card after the completion of the upfront surveys and two days of survey. The drawing of the three prizes had the goal of incenting continued participation through the 14-day survey period and respondents received one entry for each day they uploaded data and completed the associated questionnaires.

Recruitment of respondents and their participation in the survey had the following life cycle:

1. Mailing of invitational material
2. Invitational phone call
3. Initial visit for equipment delivery and training
4. Assistance during the course of the survey
5. Exit visit for equipment retrieval and incentive delivery

Advance Material

Packages with advance letters from University of Illinois at Chicago (UIC) and from the Illinois Department of Transportation (IDOT) were mailed along with an illustrated explana-

tory brochure. The letter from UIC explained the purpose and nature of the survey, and the incentives for participation. The brochure intended to provide information on the steps involved in survey completion and to attract people's interest with images of the equipment and website. Both of these items displayed UIC contact information for the clarification of doubts concerning the survey and its legitimacy. The advance letter from IDOT was included to increase the credibility of the survey in the eyes of the general public, as IDOT is an institution that is able to act upon the survey results. Past study has shown that this letter has a significant positive effect on response rate (Mohammadian et al., 2009).

Recruitment Interview

Two to seven days after the expected receipt date of the mail packages, possible respondents were contacted via telephone to assess their interest in participating in the survey, as well as the interest of any other household members. Six phone calls were tried before the household was classified as non-responsive. When the household was reached and one or more persons in the household decided to participate, an initial visit for equipment delivery and training was scheduled at the time and location of the choice of participants. The survey assistant who made this first contact remained as a unique contact person for that participant for the rest of survey, in an approach similar to that of the tailored interviewer (19). Most participants chose to have the meetings at their residence, especially the elderly. However, some preferred to meet assistants at public libraries, at their offices during lunch break or at coffee shops.

Respondents Training, Assistance and Follow-Up

During the initial visit, survey assistants: registered and trained respondents; assisted the completion of the socio-demographic, routine activities and frequently visited locations surveys; and performed the entry of the initial schedule for the planning day. Assistants then guided respondents through the completion of an example log, practicing uploading a log, correcting eventual errors on the automatically detected activity-travel pattern and clarifying any possible doubt on the meaning of survey questions. A set of illustrated, step-by-step instructions on how to use the survey equipment and website were given to respondents. The survey database was manually checked daily to ensure that respondents were completing the survey accordingly. When it was identified that a respondent had not completed the survey for more than two consecutive days, the personal survey assistant would immediately re-contact this individual to remind him/her about the survey and to make sure that no technical problem or difficulty was being faced. It was important to ensure that respondents did not accumulate more than around three days of activities and trips before submitting their answered questionnaires because otherwise the number of questionnaires to be responded to would become overwhelming and discouraging.

3.3.4.6 Data analysis

Forty-nine percent of respondents were seniors - 65 years-old or over - and the other 51% were between 18 and 65 years-old. Data on 2,160 trips and 2,131 activities was collected from the seniors and on 2,397 trips and 2,371 activities from the younger respondents, totaling 4,557 trips and 4,502 activities. The trip rate was 4.3 trips per person per day, which indicates an

above average number of trips when compared to the reference trip rate for personal travel suggested in (Stopher et al., 2008b), 3.4 trips per person per day. This result is consistent with the finding of previous studies, which demonstrate that GPS surveys have an improved ability to capture trips that are frequently under-reported in other types of survey. The non-mobility rate was 9.35%, within in the range suggested as accurate in (Madre and Brg, 2007). This result is also similar to other long duration surveys such as the seven-day German Mobility Panel and the six-week Mobidrive, which are considered to have accurate non-mobility levels (Madre and Brg, 2007).

A deeper look at the survey design, methodology, execution and analysis can be found in the associated papers (Auld et al., 2009; Frignani et al., 2010a; Frignani et al., 2010b).

3.3.5 Summary

Table I summarizes the different characteristics of the data sources listed above. As this table shows, the more time consuming planning surveys (CHASE and UTRACS) tend to be considerably smaller than the basic activity surveys, but they also offer longer survey periods providing better examples of multi-day activity patterns. This table also illustrates the need for the UTRACS survey in terms of the GPS capture alone. In addition to this GPS data, designing and executing a survey focused on the goals of this work also allowed a more detailed analysis of response time and projected savings than that captured in any other study.

3.4 Data analysis

This section explains the different ways the data used in this work were analyzed to meet the needs of the research goals. One of the challenges of hybridizing travel survey techniques with

TABLE I

DATA SET CHARACTERISTICS

Data set	**Activity Information**	**Planning Information**	**GPS Trace**	**Time Period**	**Households**
CHASE 2002-03	X	X		7 days	271
Atlanta 2001-02	X			2 days	8,069
Anchorage 2002	X			1 day	1,293
UTRACS 2009-10	X	X	X	14 days	100

real-time analysis and projections is determining the type of model needed and the necessary data to achieve the goals of this study. Described first is the model used to represent the activity patterns of an individual. This is followed by an explanation of how the dimensions of an activity based travel survey are reduced to allow for quick processing while still capturing enough detail to make meaningful predictions. Finally, the method for integrating GPS data into this model is specified.

3.4.1 Traveler history

One of the common ways activity based travel surveys are collected is by the order of activities and travel related to getting to the location of each activity. Thus, when examined by the order of activities the observations describe a chain of activities and travel. To model the activity patterns of an individual we formalize this be stating the patterns of an individual can be represented by a sequence of activity and travel events. Thus, the data is structured such that each activity or trip corresponds to an event in the person's activity pattern stream,

with a set of attributes corresponding to that pair of events that capture information about the activity and travel which we have referred to as the *traveler context* represented as TC. This model of a traveler's history is a sequence of descriptions of events or a sequence of traveler contexts. Since the traveler's history represents the sequence of these events from the time observation began until the present, this represents a continually growing sequence or stream of activity patterns. Formally, each activity event A contains a set of attributes about that activity $\{a_0 \dots a_l\}$ which is followed by a travel event B characterized by the set of attributes $\{b_0 \dots b_m\}$. By having an attribute of each of these events be the associated duration of the event, the events can be represented as a discrete sequence of sets rather than a continuous stream. This simplification greatly reduces the complexity of the problem without impacting the ability to represent that activity patterns that occur. Thus, the model of the observed activity patterns of an individual can be described as:

$$\left\langle \mathsf{TC}\left(\mathsf{A}\{a_{0,t_0},\dots,a_{l,t_0}\}\,\mathsf{B}\{b_{0,t_0},\dots,b_{m,t_0}\}\right)_{t_0},\dots,\mathsf{TC}\left(\mathsf{A}\{a_{0,t_n},\dots,a_{l,t_n}\}\,\mathsf{B}\{b_{0,t_n},\dots,b_{m,t_n}\}\right)_{t_n}\right\rangle$$

where t_0 represents time step where observation began of the individual and t_n represents the traveler context at the current time. Thus, projecting the traveler context x steps in the future would be t_{n+x}. This model is used for the remainder of this work.

3.4.2 Dimension reduction

For most applications, some form of dimension reduction is beneficial to reduce the complexity of the problem. For applications that are further constrained by needing to adapt to

changes quickly, this is especially true. With this in mind, this section explains the how and why of the dimension reduction in this work.

The first way the fields were simplified was discretization of the fields being examined. This was done for two reasons: to simplify the problem; and to generate a model that was more understandable while also allowing it to be easily compared across travelers. The second simplification was consolidation of existing discrete values that represented a range of values into larger buckets to increase the number of instances within each bucket to a meaningful size.

Finally, within this work there are two distinct types of learning related to the activity pattern of the individual. The first focuses on learning the individual's patterns based on their own history. The second is inferring likely patterns of the individual based on similar patterns of others or pattern transference. As might be expected, the data needed for these two tasks is also different.

3.4.2.1 Patterns of the individual

For learning the patterns of the individual, it is fairly intuitive that demographic information of the traveler is irrelevant since it would on a short-term scale remain constant. On a macro scale, this may not be true since the number of children could increase, or an additional job could significantly increase the income of a household. However, as described in more detail in Section 4, the learning of the individual adapts over time with more recent observations replacing distant ones. As a result, adaptation to these changes will naturally occur as time passes allowing this information to be treated as a constant.

The second part of the dimension reduction for learning the patterns of the individual relates to selecting which fields are of interest, and once those fields have been selected determining which additional fields (if any) should be added to improve predictions related to the desired fields. The validity of using this type of activity data to create traditional multi-user tour based scheduling models has been demonstrated in recent work (Doherty and Mohammadian, 2007). For this paper we selected a subset of 10 of these attributes described in Section 3.3.1 out of the 200 or more attributes typically collected in activity based surveys.

3.4.2.2 Pattern transference

The technique of pattern transference involves using the patterns of other travelers, to improve the patterns of the individual we are trying to predict for. There are multiple ways this could be accomplished, but essentially it comes down to determining when patterns are relevant to the target traveler. Unlike learning the patterns of an individual, for trying to compare other traveler patterns to improve the patterns of the individual, the attributes of the individual such as their demographic data might be highly relevant. However it is important to note the cost in terms of user burden of adding additional attributes for this purpose. If an attribute is added outside of the desired attributes described above, this would mean an ongoing burden for the traveler, however if only demographic data is added this would be essentially, be a onetime cost. As a result, the additional attributes collected for this task, explained in detail in Chapter 4, are demographic attributes.

3.4.3 Processing GPS traces

As indicated in Section 1.6, this work is primarily interested in activity locations and trip characteristics rather than the actual route of the traveler. This is important as it allows the amount of GPS information stored long term to be greatly reduced. Specifically, after the GPS trace has been processed and the activity locations, timing and trip characteristics have been extracted, the GPS trace itself is no longer needed. While keeping the GPS trace could benefit other techniques that focus on route selection, retaining them is not necessary post processing for the algorithms and techniques introduced here.

As mentioned in previous sections, to simplify the real-time learning and prediction tasks, the GPS components are discretized as well. For traits like timing and trip characteristics, this is fairly straightforward. For start, end, and duration these can be discretized by increasing the count of the appropriate bucket that contains that range of time (i.e. 2-4 pm.) For location, this is a bit more complex as we want to capture a discretized version for learning and prediction that is also flexible enough for a non-exact GPS match when the location is revisited. To accomplish this, a location id is given to activity locations detected during the GPS processing. Each of these location ids are then associated with a corresponding GPS point. By itself, this would mean a huge number of location ids that did not duplicate would be collected with little meaning to a learning component that focused on discrete patterns. To address this when new activity locations are identified from the GPS log, a comparison is made to existing locations and merged with similar previous locations resulting in a much smaller set of locations. Thus, a location id gets associated with a location circle rather than a single point. This is important

as it keeps the number of locations that must be stored significantly smaller, while also making the learning algorithms job and readability of the model easier. The technical details of how this is accomplished are described in Chapter 4.

3.4.4 Evaluation across studies

One of the complexities of examining traveler pattern transference is that the exact features captured by various studies can be different. For example in the three data sets used for testing, while certain attributes were the same, a large number of them differed. Some of the common differences were different names, for example Anchorage referred to an arrival time as "arrive" while Atlanta used "arrtm," despite these two surveys being conducted by the same company, NuStats. This process of mapping the same concepts across data sets is known as ontology alignment (Gruber, 1995). While the alignment mapping across the surveys can be easy, as described in the previous example, other attribute differences can be more complex where the mapping is more fuzzy. For example the Anchorage survey has a salary range of "$70,000 to $79,999," while the Atlanta survey has "$60,000 to $74,999" making the matching more gray than black and white.

Due to these differences between these data sources, creating an alignment between them was necessary to allow patterns to be meaningfully transferred between the different areas captured by the surveys. The alignment used for demonstrating transference was created by hand mapping the fields used between the data sets; however there is promise of creating (or at least starting) this type of mapping automatically as shown in Section 4.5. It is important to note that while it may seem advantageous to create a hand mapping of the alignment in such

a way as to skew the results in this studies favorably, in fact the opposite is true. This is due to a less precise mapping degrading the value of the patterns learned and thus reducing the accuracy of the predictions.

3.4.5 Testing methodology

For all experiments conducted in this work a 10 times cross-folding or hold-out methodology was used to ensure the validity of the results. A k times cross-folding methodology (where $k \geq 2$) refers to randomly splitting the instances of the data set into k sets of data of approximately equal size. Tests are then carried out using $k-1$ sets as the training set and the remaining set (the hold-out set) as the test set. This process is repeated until each set has been used as a test set and the average of the k different runs is reported. The only exception to this method is in the transferability testing where one city was used for the training set and a second city was used as the test data set. However, the results reported for a city being both the training city and the test city, once again used a 10 time cross-folding methodology as stated in 4.4.2.2.

All tests (except for the activity pattern transference where one city was used for training and another city was used for testing) used a 10 times cross-folding methodology. Cross-folding is the more common name for hold-out testing within the data mining community. Applying a 10 times cross-folding methodology means that the data is randomly split into 10 segments of as equal size as possible, and 9 of the segments are used as the training sample and the remaining segment is used as the test data. This is repeated so that each of the 10 segments is used for the test set and the average of the 10 runs is reported.

3.4.6 Metrics

For measuring prediction performance, we use the information retrieval metrics of precision and recall (Cleverdon, 1970). The basic definition of recall and precision can be written as:

$$precision = \frac{\#\ true\ positives}{(\#\ true\ positives + \#\ false\ positives)}$$

$$recall = \frac{\#\ true\ positives}{(\#\ true\ positives + \#\ false\ negatives)}$$

For the purpose of this study, since we are primarily interested in the correctness of the attribute value (if the attribute appeared at all). Thus *# true positives* is the number of attribute values predicted correctly; *# false positives* is the number of attribute values incorrectly predicted where the attribute did appear in the time step, and *# false negatives* is the number of attributes no value was predicted for, but some value for the attribute appeared in the time step. Since these two measures are often associated with a trade off of one for the other, we also examine a combined metric the F-measure (van Rijsbergen, 1979) which can be calculated as:

$$F\text{-}measure = \frac{(2 \cdot precision \cdot recall)}{precision + recall}$$

We use this metric to compare the balanced performance of the algorithms.

3.5 Summary of methodology

In this section, an overview was given of the methodology applied in this study. For this study, a quantitative research design was used to empirically demonstrate the findings of this

work. A quantitative design was chosen since it was best suited for the task of data mining and prediction of known variables. However, in addition to quantitative measures, qualitative comparisons are also used for comparing readability of the models generated.

To approach the problem of learning activity patterns of individuals, a subset of attributes typically used in urban planning activity surveys was used. Specifically the attributes selected capture activity, location, mode of transportation, arrival time, departure time, duration, sequential order, and planning flexibility information of the activity and travel of the person. These attributes were picked because while they capture the information of interest in modeling the traveler, they also have been shown to effectively predict travel and activity patterns of types of individuals in planning studies (Doherty and Mohammadian, 2007).

The traits of interest were discretized and the series of activities and associated travel of the participant became a series of sets representing the events or traveler context throughout the day. The findings of this study were demonstrated using three transportation surveys collected outside this study in addition to a survey conducted as part of this work. Due to differences between the different studies, the data captured in the studies had to be aligned by hand such that the patterns identified in one survey could be meaningfully examined in relation to the other studies. To fit this same model the GPS trace that was collected as part of this work was processed to automatically identify discretized versions of the attributes matching the study's goals.

Thus, the goal of this study was given a series of traveler contexts gathered through active and passive means, predict the next set or sets of contexts or missing attributes of interest in

previous contexts. Information retrieval measures were then used to quantitatively analyze the value of the predictions of the model created. To demonstrate the soundness of this work; the primary goal was broken down into a number of sub-problems. First, the internal validity of being able to predict the activity patterns of a traveler with "perfect" information is shown. Next, techniques are introduced and analyzed for reducing the burden on the traveler for collecting data. This is followed by presenting algorithms and techniques that can further reduce the amount of data that needs to be collected with limited impact on the quality of predictions. Afterward, a demonstration of how activity patterns might be transferred from other locations to improve predictions is shown. Finally, a combined framework of how all of these concepts might be brought together for quickly learning the activity patterns of an individual is presented.

CHAPTER 4

RESULTS AND CONCEPTUAL MODEL

4.1 Traveler context prediction

This section is based on previously published work (Williams et al., 2008).

Recent work has focused on creating models for generating traveler behavior for micro simulations. With the increase in hand held computers and GPS devices, there is likely to be an increasing demand for extending this idea to predicting an individual's future travel plans for devices such as a smart traveler's assistants. In this section, we introduce a technique based on sequential data mining for predicting multiple aspects of an individual's next activity using a combination of user history and their similarity to other travelers. The proposed technique is empirically shown to perform better than more traditional approaches to this problem.

4.1.1 Introduction

Modeling and predicting travel patterns has been a source of interest and a challenge for regional planners for a number of decades. In the past, data available for these studies has by in large been limited to aggregate traffic information and paper based travel surveys. With the increase in prevalence of hand held computers and GPS devices, the detail and types of information that can be captured for travel histories are becoming far richer than ever before.

While these data sources share many characteristics with other studied travel activity logs, the additional information on decision processes and opportunities for further enhancement

with GIS offer some unique challenges as well as opportunities in behavior prediction. In this section, we focus on techniques for extracting meaning from these augmented data sources. One of the challenges associated with understanding traveler behavior is its strong dependency on the transportation network. The strength of this relationship makes it nearly impossible to address behavior without also considering the overall *traveler context* in terms of location, mode of transit, what is being done, and the time of the activity to name a few aspects. A better understanding of how this context affects future behavior is likely to not only improve predictive models, but also provide additional insight into how choices are made among alternatives.

In this section, we explore the challenge of traveler prediction not as activity or location prediction, but as context prediction. Specifically we examine the problem of given a sequence of the traveler's prior context how well can the context of their next activity be predicted. Rather than focusing on prediction of common activity sequences based on the history of many travelers, we shift the paradigm to predicting the travel context of an individual based on their own travel history and the history of others. This subtle shift is to facilitate moving travel prediction from the realm of urban planning to enable personalized smart traveler assistant devices. The principle behind these devices are to provide intelligent services like multi-modal travel planning to support a user's travel needs (Dillenburg et al., 2002; Torrens et al., 2004). With this type of application, understanding the context of a traveler can help identify what type of information the user is interested in and help narrow potential options. Consider a scenario where a user wants to make a trip across town mid-day. If the assistant knew the

user's likely mode of transit and accessibility thresholds, the system might be able to evaluate and suggest one transit route over another.

We propose a novel approach to predicting individual traveler behavior and associated context from a stream of their prior traveler context. The proposed approach examines each step in the history and the predicted next step of the traveler as a set of attributes or characteristics describing the context of that step. As a whole, these steps form a sequence of sets describing the traveler's context progressing through time, or an enhanced activity sequence. We introduce a technique for mining sequential associative rules from individual activity sequences and augment recall and thus overall performance by incorporating the patterns of other travelers through the traveler's context.

4.1.2 Mining Multi-Variate Streams

In this section we introduce a method for mining temporal stream data based on constrained associative sequence mining introduced in Agrawal and Srikant (Agrawal and Srikant, 1995). Our contribution is an extension of these techniques for next step multi-variate prediction within a single discrete temporal stream. In this context, the multi-variate prediction refers to the prediction of the multiple variables or characteristics associated with the next traveler context simultaneously, as opposed to trying to classify each characteristic individually.

Association mining and sequential association mining has traditionally been applied in the business world for analyzing purchase patterns through market basket analysis. Since this technique was introduced in 1993 (Agrawal et al., 1993), associative pattern analysis has remained widely popular due in part to the readability of the found patterns and the intuitiveness of the

strength of the rule (Hipp et al., 2000). The output of this type of sequential associative mining is a rule of the form:

$$\langle\{\mathsf{AB}\}\{\mathsf{C}\}\rangle \Rightarrow \{\mathsf{DE}\}, \text{with support} = 25\%, \text{and}$$
$$\text{confidence} = 80\%$$

This rule can be read as: if the set {AB} is followed by the set {C} this implies the set {DE} will appear in the future. The strength of this pattern is that this full sequence, $\langle\{\mathsf{AB}\}\{\mathsf{C}\}\{\mathsf{DE}\}\rangle$, occurred in 25% of the training examples (support); and for all training examples where $\langle\{\mathsf{AB}\}\{\mathsf{C}\}\rangle$ appeared, the full sequence $\langle\{\mathsf{AB}\}\{\mathsf{C}\}\{\mathsf{DE}\}\rangle$ appeared 80% of the time (confidence). Thus, as this example shows, unlike some other model-based techniques for example neural networks, the model and patterns identified by this technique can be further analyzed and explained if desired. In addition, because the sequential patterns identified are only order dependent, meaning additional sets of items could occur between the identified sets without consequence as long as the sets are in that order, the technique is also more robust to noise in terms of still being able to identify the underlying pattern. From a traveler prediction perspective, this feature is likely to help in identifying the underlying patterns despite sporadic activities that may not be part of the regular routine of a traveler's day.

Identifying patterns in traditional associative mining relies on multiple training sets for its primary constraint *support* (Agrawal et al., 1993). With associative sequence mining, there is a similar dependency on multiple training sequences (Agrawal and Srikant, 1995). The implication of this when applied to the context of transportation is that for a travel or activity

pattern to be significant, the pattern must be present across multiple travelers. While this constraint is likely a good guideline for predicting traveler patterns in general, if the goal is to predict the travel pattern of an individual, then patterns that are unique to that individual are likely significant for predicting future behavior of that individual even if they have little predictive value for the set of all travelers as a whole. In addition, these techniques are not well suited to lengthy sequences, as the distance between sets within a sequence is not accounted for. Applying this logic within the context of transportation, this would be equivalent to saying the likelihood of an event occurring is just as dependent on an activity that occurred 4 days ago as it is on the previous activity. While such relationships may exist, it seems reasonable to assert that activities that occurred in the traveler's recent history are in general more likely to be better predictors of the immediate next activity.

Based on these perceived weaknesses of existing associative sequential mining techniques when applied to individual traveler prediction; we propose an alternative to the traditional constraints which we assert is better suited to this type of prediction problem. The purpose of the proposed technique is to identify frequent sequences within multi-variate temporal data, such as the many different characteristics describing the activity of a traveler, that can be used when patterns within a single stream are meaningful, and is more suited for lengthy sequences. Below we present an approach to constraining the problem of frequent sequence mining to allow existing associative sequential rule mining algorithms to work in this context.

4.1.2.1 Revised Constraint Definitions

An important difference between this new problem formulation and the typical data used with sequential association rule mining is the number of sequences. In traditional sequential association rule mining there are many sequences, each representing an instance of a specific occurrence of a sequence of item sets. As such, traditionally how frequently an item set or sequence of item sets occurs is defined in terms of the number of training sequences the set(s) occurs in, more formally support count of X is defined as:

$$X.\text{count} = \frac{\text{\# of training sequences containing the item set X}}{\text{total \# of training sequences}},$$

and the definition of the support of a sequence of an arbitrary number of item sets $X\ Y \ldots Z$ is defined as:

$$\text{support}(\langle\{X\}\{Y\}\{\ldots\}\{Z\}\rangle) = \frac{(X_i \cup Y_j \cup \ldots \cup Z_k).\text{count}}{\text{total \# of training sequences}},$$

where $i < j < k$ (number of times the item sets appear in that same order). By contrast, since we want to identify patterns in a single training sequence, we are interested in how frequently a set of items occurs in terms of the total number of item sets in the user's sequence. Thus, we define the count of an item set as:

$$X.\text{count}' = \frac{\text{\# of item sets containing(the items of set X)}}{\text{total \# of item sets in the training sequence}},$$

and the definition of the support of an arbitrarily long sequence of item sets, $X\ Y \ldots Z$, is thus also defined in terms of item sets, but additional consideration is taken since we are trying to capture support of sequences of item sets within a single sequence. Thus, without additional constraints any item anywhere in the sequence after the first item set could be considered support for the sequence regardless of their distance apart in the sequence. We thus want to constrain the number of item sets considered after the first item set in the sequence in question. To address this we use the concept of a **support window**. A support window specifies the item sets after the first item set to consider when calculating support. Thus when calculating the support of any potential frequent sub-sequence within the original sequence, there is a support window of some specified length w that slides to the first item set considered. Using this concept we define the support of a sequence of an arbitrary number of item sets $X\ Y \ldots Z$ to be defined as:

$$\text{support}'(\langle\{X\}\{Y\}\{\ldots\}\{Z\}\rangle) = \frac{(X_i \cup Y_j \cup \ldots \cup Z_k).\text{count}'}{\text{total \# of support windows in the training sequence}}$$

where $i < j < k \leq (i + w)$. Thus the number of times the sequence occurs within the window constraints of all possible windows in the sequence. All other common multiple support sequential mining constraints designed to help limit the search space such as λ (relative minimum support), support difference constraint, and minimum support threshold are calculated with respect to this new support measure (Liu et al., 1999). Likewise, when rules are mined

from the extracted frequent item sets, confidence is defined in terms of this definition of support.

The results reported in this section were obtained by applying these constraints to the Generalized Sequential Patterns (GSP) algorithm (Srikant and Agrawal, 1996), although it should be noted that any multiple support sequential association rule mining algorithm such as PrefixSpan (Pei et al., 2004) could be used with this technique.

4.1.3 Related Work

Techniques for constraining and mining sequence association rules have been extensively studied since associative mining and later sequential mining techniques were introduced by Agrawal *et al.* (Agrawal et al., 1993; Agrawal and Srikant, 1995). The sequential mining techniques introduced in that work also explored a similar concept to the sliding windows discussed here, but this idea was only explored in terms of defining support between sequences rather than as a mechanism for discovering patterns within a single sequence (Agrawal and Srikant, 1995). Zaki introduced length, width, and maximum gap constraints to reduce data mining time. In Harms *et al.* and Harms and Deogun an algorithm was introduced for mining frequent episodes from multiple sequences using time lag constraints and separating antecedent and consequent constraints (Harms et al., 2002; Harms and Deogun, 2004). All of these works have focused on extracting and applying the associative rules uniformly across sequence series in prediction. In addition, all of these works tend to focus on predicting a single next item in a given sequence. Other work has examined set prediction as constrained sequence mining, but focused on traditional sequential association rules (Garofalakis et al., 1999). Work on label

sequential rules were introduced and examined in the context of text mining for their benefits at being able to discern the context and applicability of rules (Liu et al., 2005; Jindal and Liu, 2006).

Much of the related transportation specific research falls into two general categories: micro-simulation, and individual travel prediction. In the area of micro-simulation, related work has primarily focused on using activity survey data for generating simulated activity schedules or verifying simulations (Pribyl and Goulias, 2005; Lee and McNally, 2003). Recent work has examined using mental maps and cognitive learning for improving choice models through observations during micro-simulations (Arentze and Timmermans, 2005). However all of this work has focused on simulating the behavior of artificial travelers. Other work has focused on predicting next location of individuals based on GPS traces (Ashbrook and Starner, 2003). Liao et al. extended this idea and examined this problem as an unlabeled activity model for predicting the next location (Liao et al., 2007). This study differs from these in that like the GPS based techniques, it focuses on individual behavior, but our goal is more similar to micro-simulation models, in that we are interested in modeling a more feature rich set describing the reason for the behavior beyond the locations.

4.1.4 Experiments and Discussion

4.1.4.1 Methods

Data

In this section we use a dataset derived from the Computerized Household Activity Scheduling Elicitor (CHASE) survey to evaluate the proposed constrained rule mining problem (Doherty

and Miller, 2000). CHASE was conducted in Toronto between 2002-2003 and consists of travel activity information for 426 adults in 271 households. This survey focused on observed activity and travel patterns as well as the planning process associated with these activities captured over a 7 day period for each individual. This data set is described in detail in Section 3.3.3.1.For this section we have selected a subset of these attributes (18 of approximately 200) and only activities that are adults only, current acts, good data, no driving acts, and no doing the survey acts. These 18 attributes listed below describe the portion of the traveler context examined in this section and all have discrete values:

- Activity Group
- Activity Specific Type
- Activity group (3 categories related to if tour was mandatory)
- Activity group (5 categories related to involvement of others in tour)
- Activity group (8 categories for tour analysis)
- Total number of children under care in the household that are involved
- Location id
- Location, observed (in or out-of-home)
- City
- Tour type: At-home (AH) or home-based (HB)
- Observed sequence of the AH or HB tour (1st AH tour of the day, 2nd AH tour etc)

- Observed sequence of the activity in the tour (1st act in the tour, 2nd act in the tour,...)
- Duration flexibility binary indicator
- Time flexibility indicator
- Spatial flexibility indicator
- Whom the activity is conducted with
- Whom the activity is normally conducted with
- Interpersonal flexibility

These attributes were selected as they represent a mix of information about the type of activity, the location, relative time the activity took place (# tour of the day), and the person's involved. Thus, the dataset can be thought of as a stream or sequence of events with the set of attribute values at each event being highly related. For this dataset, there are 41,312 sets of traveler context, with the average adult's traveler context sequence being just over 92 sets in length.

4.1.4.2 Experiments and Evaluation

In the results below, we also present the results of Naive Bayes and a first order Markov model for the purpose of a comparative baseline. For both of these models, a classifier was built for each of the 18 attributes potentially present in each time step. The results presented below evaluate the set produced by this group of classifiers at predicting the next attribute value set at the next step in the series. The numbers presented are the average across the 18 attributes.

To ensure the significance of our results, we have employed a 10 times cross-folding validation methodology. Since sequential data is by definition order sensitive, care must be taken to ensure order is preserved as much as possible while still producing meaningful cross-folding. For this we employed a loop technique designed to keep as much of the training data as sequential as possible while still allowing the test set to immediately follow the training data that would have been seen just prior to the test data in the actual event. To illustrate this, consider a sequence S that would be split into parts $S_1, S_2, \ldots, S_5$. The resulting test series would be:

$$\begin{aligned}
\mathsf{Train}_1 &= S_1S_2S_3S_4 \qquad \mathsf{Test}_1 = S_5,\\
\mathsf{Train}_2 &= S_5S_1S_2S_3 \qquad \mathsf{Test}_2 = S_4,\\
\mathsf{Train}_3 &= S_4S_5S_1S_2 \qquad \mathsf{Test}_3 = S_3,\\
&\ldots
\end{aligned}$$

All results shown are the average of the 10 runs.

In the first set of experiments, we examine the impact of minimum support on predictive quality using a support window of 3 and a window size of 1 for predictions. Figure 1 and Figure 2 depict the effect minimum support has on precision and recall respectively with the additional support window constraint. As these figures show, like traditional association rules, the higher the minimum support, the more precise the rules, but there is a tradeoff of lower recall. Figure 3 displays the F-measure of these same models and show that the lower minimum support of 10% combined with a confidence threshold of 66% provides the best balance of precision and recall.

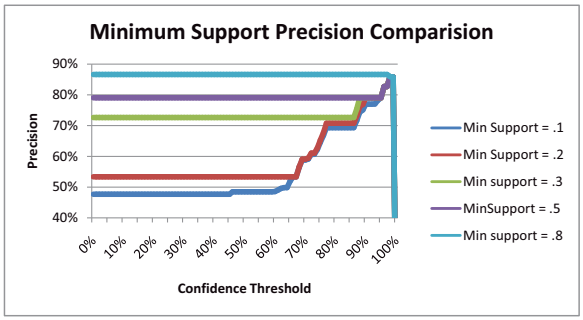

Figure 1. Comparison of precision for various minimum support.

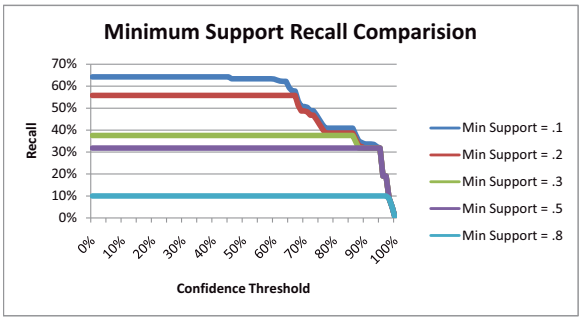

Figure 2. Comparison of recall for various minimum supports.

In the next set of experiments we compare the support window constrained algorithm, with two common algorithms as baselines. For the support window constrained algorithm we selected the best mix of support window, prediction window, and minimum support constraints based on F-measure for comparison; specifically a support window of 3, a prediction window of 2, and minimum support of 10%. Figure 4 and Figure 5 depict how the proposed algorithm compares with each of these baselines with respect to precision and recall. As Figure 4 shows the proposed

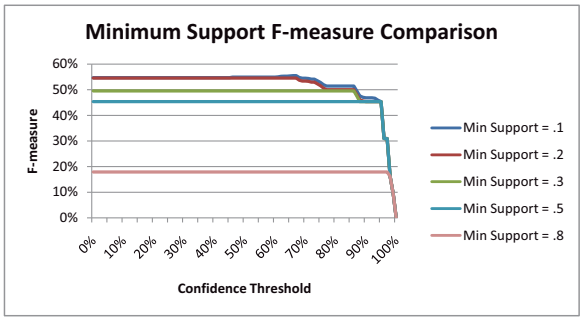

Figure 3. Comparison of F-measure for various minimum support.

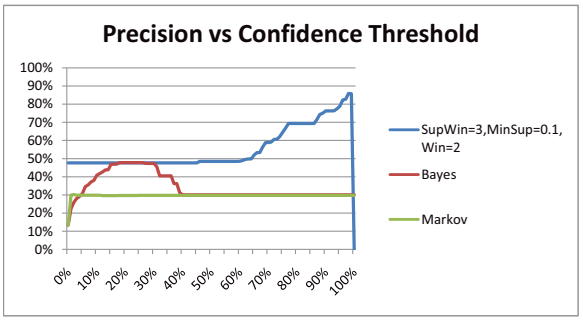

Figure 4. Comparison of precision for algorithms.

algorithm is as precise as either of the baselines across the entire confidence threshold range. A closer look at the recall shows that while Bayes has a higher recall at low confidence thresholds, the proposed algorithm has a better F-measure for the entire range as seen in Figure 6.

In the last set of experiments, the prediction performance of a selection of the attributes is presented in Figure 7. This chart contains a comparison of how well each attribute is predicted beyond the overall averaged predictions. As these results demonstrate, the location selection is

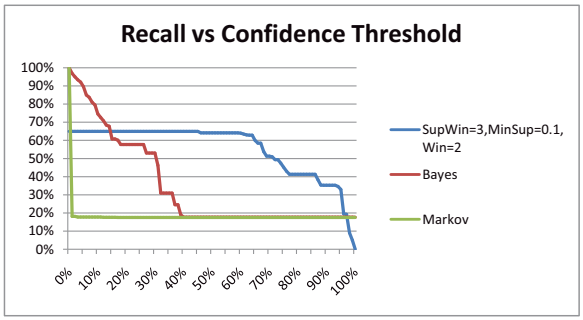

Figure 5. Comparison of recall for algorithms.

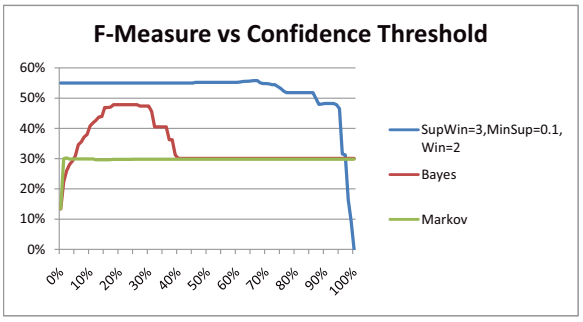

Figure 6. Comparison of F-measure for algorithms.

one of the easiest attributes to predict as it likely best captures traveler's preference to frequent locations as part of their routine. Another interesting aspect of these results is that while the locations are fairly predictable the activity is more difficult to predict. This indicates that people often do several different activities at the same location at different times, which demonstrates location prediction alone is insufficient in determining the reason for travel. The missing data is interesting as it reflects which attributes are likely to get non-responses from users. As the

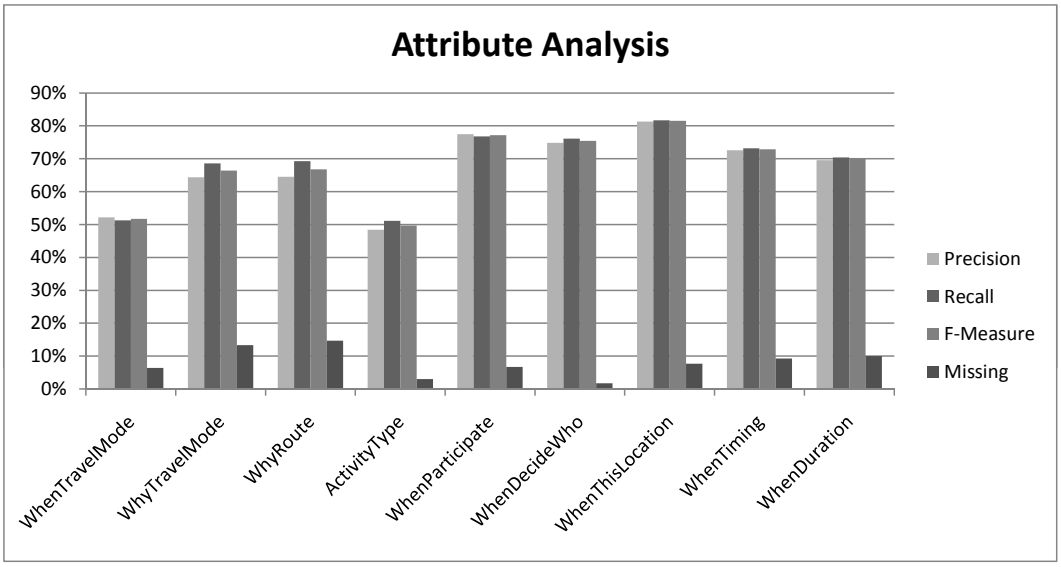

Figure 7. Comparison of attribute prediction performance.

percentage-missing shows, while the activity and location are commonly entered, capturing when items are planned and the reason for selecting their travel are harder for participants to recall perhaps due to people's tendency to follow their routine rather than consciously thinking through each decision.

4.1.4.3 Discussion

In this section we introduced a technique for constraining sequential associative mining algorithms so that individual streams of activity sequences can be mined effectively. As our results demonstrate, this technique shows significant promise for accurately predicting next step context compared to Bayesian and Markov approaches. This advantage is likely due to the strength of relationship within the predicted set being better captured by associative mining.

For simplicity in the experiments above, the support window was specified by a fixed number of item sets; however it is easy to envision more complex notions of this same concept specific to transportation. For example, a more meaningful boundary in transportation may be within a day or within a tour. The benefit of a more complex transportation specific notion such as a tour for a support window is explored in Section 4.4. As shown, an additional benefit of this technique is that the predictive model is much more transparent than techniques such as neural networks, allowing additional insight to be gained from the models created.

4.2 Data collection and reducing user burden

This section is based on previously published work (Auld et al., 2009; Frignani et al., 2010a).

One of the challenges in learning the behavior of a user for mobile applications is the balance between collecting enough data to learn the person's patterns and over burdening the user with the amount of interaction required. When discussing the effort required by the user we refer to two types data collection *active* and *passive*. In this context *active* data collection refers to data that is collected by the traveler having to manually enter information specifically for purpose of learning their behavior. *Passive* data collection refers to data that can be collected from other sources without user interaction. The previous sections have validated the use of activity context to project future activity contexts, given data actively provided by the traveler. While this type of actively collected data has been proven to provide good predictions, it also represents a significant burden on the traveler.

A primary goal of this work was to gain the benefits of traveler activity pattern modeling without the burden on the traveler that is usually associated with the amount of data needed

from the person. The approach proposed here is based on leveraging passive data sources such as GPS traces and learning methods to reduce the data entry requirements of the user. As such, a data source was needed to examine the benefit of passive data analysis as well as some method of measuring the reduction in burden on the user. To address this, as described in Section 3.3.4, information on the detected location accuracy as well as time required by the user for questions was recorded. In this section, a GPS processing method is introduced and evaluated for its ability to passively detect significant locations. This is followed by a discussion of the ways techniques discussed in this section were applied to the UTRACS survey, how they performed, and finally a discussion of how the burden may be further reduced.

4.2.1 Detection of significant locations using GPS

One of the key themes throughout this work has been exploring ways to reduce the effort required by the user while still being able to make reasonable predictions. One of the primary ways of accomplishing this is through leveraging passive data sources such as GPS in helping establish context without user input. Within this section a new approach to GPS processing is introduced, followed by an evaluation of the accuracy of the approach.

4.2.1.1 Introduction and motivation

Many studies have examined ways GPS may be used to aid in detecting significant locations, activities, and travel characteristics (Rudloff and Ray, 2010; Liao, 2006). One of the common characteristics of these approaches is detecting these aspects or its modeling takes place during post-processing. For example, Ashbrook and Starner's approach was based on k-means clustering over a history of the travel of the user to detect most frequented locations (Ashbrook and

Starner, 2003). Other work by Flamm *et al.* used spatial density algorithms to identify significant locations (Flamm et al., 2007). By contrast here we introduce an approach that detects significant locations, matches the locations with previously detected location, and detects the start/stop of trips as the GPS data is recorded.

The basic method introduced here, like similar work, is to use a threshold based evaluation on time and distance. The key difference between the approach presented here and prior work is to use the aggregation of points rather than an ongoing comparisons of individual points. This contribution is significant as stops can be identified, matched with prior locations, and the resumption of travel detected all on the fly rather than waiting until post-processing. The significance of this is that the actual context and better identification of prior context can be determined on an ongoing basis without user input. By gathering ongoing location and trip information unobtrusively, a prediction model based on prior history can be much more accurately built with far less effort than previously required by the user.

4.2.1.2 Problem statement and approach

Given an ongoing stream of GPS points from a multi-modal traveler, detect incrementally:

- Significant stops/locations as defined by a specified distance and time threshold
- Match these with previously identified locations and update location if necessary, and
- Detect the resumption of travel

The approach taken to accomplish this is to use the incoming GPS information to estimate the likely constraints of a potentially significant location. The estimated bounds of the potential

location are approximated as being represented by a circle around the point with the diameter of the potential location being the distance threshold. Thus as new points are recorded, as long as they are within the estimated location circle no other additional comparisons need be made against all other points of the trace. The determination of if a current location is significant is then made based on if the difference in time from the last recorded point to the first recorded point of the circle is greater than the minimum time threshold. As these significant locations are identified, they are merged with previously identified locations. Other approaches to this problem have been used in geomatics such as map matching, however the vast majority of these techniques rely on matching GPS traces to known routes or locations (Quddus et al., 2003; Blewitt and Taylor, 2002). As a result, one of the limitations is very detailed maps are needed of the area in question. Applying these techniques to the context of this research would mean at a fine grain locations would need to be labeled in addition since this work focuses on multi-modal travel paths which are not roadways would also need to be considered which are not readily available for most areas. Thus, while there is significant potential for these techniques being able to provide a richer source additional data from passive sources, at the time of this work the detail and coverage of these maps was determined to be insufficient for a detailed study within this work.

4.2.1.3 Data preparation

One of the challenges of many GPS loggers at the time of this study that also plagued the devices used in the UTRACS study was the problem of invalid points being recorded by the loggers. The primarily cause of these errors are due to interrupted of satellite visibility. Some of

the causes of this are due to the well known urban canyon problem where there is only limited satellite visibility preventing a good location fix or due complete signal loss as might occur when entering a building. To address this problem a filter was applied so that only "valid" GPS points with sufficient position information were considered. Specifically only points where at least 4 satellites were visible and the horizontal dilution of precision (HDOP) is less than 5 (Stopher et al., 2005b). By adding this criteria the amount of obviously incorrect points was cut down considerably which is discussed further Auld *et al.* (Auld et al., 2009).

4.2.1.4 Significant location determination

This section describes the design of the algorithm used to determine significant locations. To determine a significant location a distance threshold $\mathtt{d}_{\text{threshold}}$ and time threshold $\mathtt{t}_{\text{threshold}}$ are used. The problem being addressed is how to find a set of points that are close enough for the set period of time. While this could be accomplished by repeatedly checking all previous points from the current time $\mathtt{t}$ back to $\mathtt{t} - \mathtt{t}_{\text{threshold}}$ until all points in the set are within the specified distance from each other, this is highly inefficient. Instead an approach is used that is based on comparing points to the estimated current significant location. The basic approach is to estimate the circle enclosing the potential significant location, then new points can be evaluated based on if they are enclosed in the circle rather than needing to check against a large number of previous points. By taking this approach significant locations can be detected as points incrementally come in rather than relying on an analysis of all points as is required with current algorithms that focus on significant location identification during post-processing.

The basic method applied is on the first point read, estimate the current significant location as a circle centered at that point with a diameter equal to $\mathtt{d}_{\text{threshold}}$. As subsequent points are read, they fall into one of two cases. Case 1, if the point falls within this existing circle it is subsequently added to the current set of points until the difference in the time between the first point and the current point exceeds $\mathtt{t}_{\text{threshold}}$ at which point the set of points is identified as a significant location. Otherwise case 2, the center of the minimum enclosing circle (MEC) of the points collected thus far becomes the new center of the circle and the diameter remains $\mathtt{d}_{\text{threshold}}$. If the current point now falls within the circle add the point to the collection and the algorithm is ready to read then next point. Otherwise check to see if the existing circle is a significant location (time meets $\mathtt{t}_{\text{threshold}}$), if it is finalize the location and note that travel has resumed. If however the previous points did not make up a significant location, go through previous points in reverse order until the previous point is farther than the threshold distance from the current point. All points prior to the point that is beyond the threshold distance and that point itself are removed from the current collection and the next point is ready to be read.

Concepts

- *Significant location:* The physical location where an activity took place.
- *Finalized Location:* A finalized location is a representation of a significant location that has been observed in the traveler's history. This location is represented by a circle with a diameter $\mathtt{d} \leq \mathtt{d}_{\text{threshold}}$. The circle of the finalized location is intended to enclose the actual significant location.

- *Assemblage Location:* One of the key concepts of this algorithm is the idea of assemblage locations. An assemblage location is a set of points $p_{t-i} \cdots p_{t-1}$ and/or finalized locations that together can represent a location. The GPS points $p_{t-i} \cdots p_{t-1}$ represents all points that are contained by the assemblage location from the first point p_{t-1} up until the previous point p_{t-1}. Thus the time span represented by the assemblage is $i - 1$. This location is represented as a circle with a center and diameter of $d_{threshold}$ such that all points and finalized locations within the assemblage are contained within that circle. Assemblage locations essentially represent an estimate of the potential significant location based on the points that have been observed. As explained in more detail below, if an assemblage has met the $t_{threshold}$ requirement and has been matched with a previous significant location, it will contain the finalized location representing that significant location.

Functions

- **FinalizeLocation**(L) takes an assemblage location with a timespan greater than or equal to $t_{threshold}$ and returns a finalized location represented by the MEC of all the elements (points and locations) of the assemblage location.
- **CompatibleLocations**(L1 , L2 , $d_{threshold}$) compares two locations and returns a location that encompasses both locations if one exists with a diameter less than or equal to the distance threshold, $d_{threshold}$, otherwise $\emptyset$ is returned. It is important to note that a location can be a single GPS entry, an assemblage location, or a finalized location, so this function can be used to determine if a new point is compatible with an existing assemblage location.

- **AdvanceLocation**(p_t,L,$d_{threshold}$) takes a new GPS point and an incompatible assemblage location which has a time span less than the time threshold. This function returns an assemblage location containing the passed point and all other points from p_t to p_{t-j} where p_{t-j-1} is the first point that has a distance between itself and p_t which is greater than $d_{threshold}$.
- **MergeWithDetected**(L, D) takes an assemblage location L and the database of finalized locations D and finds the closest finalized location C in D and if L and C are compatible add C to the elements L is representing. This means that the current location is the same as the previously visited significant location represented by C.

Algorithm

Once a significant location has been detected, the next step is to evaluate if the location matches a previously visited location. This is accomplished by first finding the closest significant location from the history. To determine if they are the same location, the minimum enclosing circle MEC of the location from the history and the collection of points from the current trace is calculated and if the diameter of the MEC is less than the distance threshold, the points are determined to be for the same location.

After a significant location has been determined, whether previously visited or not, all subsequent points that fall within the approximated circle are added to the detected location. If a new GPS point falls outside of the initial estimated bounds its inclusion in the current location is determined by whether the diameter of the minimum enclosing circle (MEC) of the existing points and the new point is less than or equal to the distance threshold. Once a

point is read which is no longer contained within the MEC of the previous points, two things occur. First, the previous location is *finalized*. Second, the subsequent points are treated as the resumption of travel.

The finalization of the location entails calculating the MEC of all compatible points from the current trace and the circle of any previously detected location(s); the new significant location represented by the MEC is added to the history of significant locations and any observations related to any previous significant locations contained by the new location are updated to the new location. The points and locations contained by the MEC are then discarded and only the MEC center and radius are kept, thus minimizing the data to be kept track of and the number of future comparisons. The the main algorithm for this is:

ProcessEntry($\mathsf{p_t}$ **,** L **,** D**,** $\mathsf{d_{threshold}}$**,** $\mathsf{t_{threshold}}$**)** which takes the current GPS entry $\mathsf{p_t}$, the current assemblage location L, the database of current finalized locations D, the distance threshold $\mathsf{d_{threshold}}$ and time threshold $\mathsf{t_{threshold}}$; and updates the location information based on the passed point and returns the resulting current assemblage location. This location could be a significant location, or an assemblage location that has yet not met the time threshold thus indicating travel.

1. (Initialization) If L is $\emptyset$ return a new assemblage location centered at the passed point $\mathsf{p_t}$
2. $\mathsf{L_{temp}} \leftarrow$ CompatibleLocations(L,$\mathsf{p_t}$,$\mathsf{d_{threshold}}$)
3. If $\mathsf{L_{temp}} = \emptyset$
    (a) If L timespan is greater than or equal to $\mathsf{t_{threshold}}$

i. $L_{final} \leftarrow$ FinalizeLocation(L)

ii. $D \leftarrow D \cup L_{final}$

iii. Return a new assemblage location centered at the passed point p_t

(b) Else return $\leftarrow$ AdvanceLocation(p_t,L,$d_{threshold}$)

4. Else if L timespan $t + 1 = t_{threshold}$

(a) $L' \leftarrow$ MergeWithDetected(L,D)

5. Return L'

4.2.1.5 Evaluation

The performance of this algorithm was evaluated using a pilot test of the UTRACS survey. The test was conducted on 5 participants with an average log of 8 days for each survey respondent. The detected significant locations were presented to the survey participant to approve the captured activity location, change the location, add additional locations, or remove the location. Thus approvals would be matched locations, additions would be missed locations and deletions incorrectly identified locations. The survey participants recorded 197 activities during the survey period.

To determine the distance and time threshold to use for the algorithm experiments, first, the distance threshold was determined by trying distances of 25m, 50m, 75m, 100m, 125m, and 150m. As Figure 8 shows the low distance thresholds suffered from missing many locations due to either the traveler moving beyond the threshold during the activity or perhaps due to slight drifting of position associated with imperfect satellite positioning. However, at 100m this seems

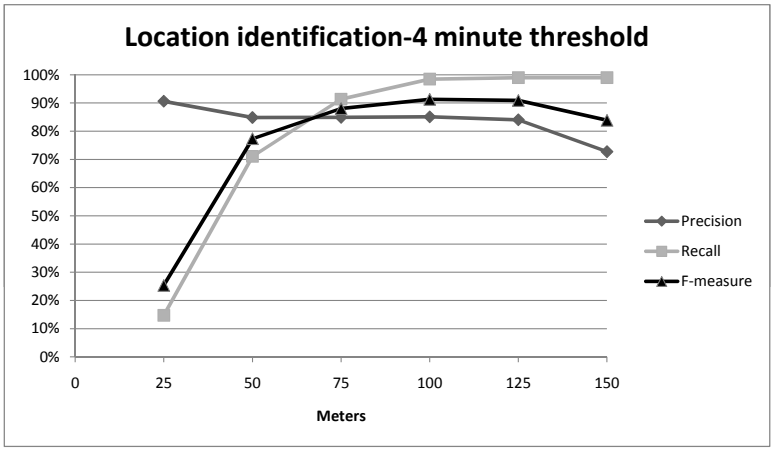

Figure 8. Location identification performance by distance threshold.

to provide a distance large enough to catch most activities while still being small enough to keep the precision high. As a result, 100m was used as the distance threshold.

The second set of tests examined how varying the time threshold affected prediction performance. Using the 100m distance threshold found in the previous set of experiments the time threshold was tested at intervals of 1, 2, 3, 4, and 5 minutes. Figure 8 demonstrates that at a low time threshold many stops due to traffic lights or simply being delayed result in many locations being identified as significant though they are not as shown in the precision. From a perspective of identifying all the significant points this remains high until it begins to significantly drop off at 5 minutes. These reflects some short stops such as pick up or drop off being missed due to the stops being quick. Based on these results the best combination of distance and time thresholds, 100m and 4 minutes, was selected for competitive comparison with other techniques.

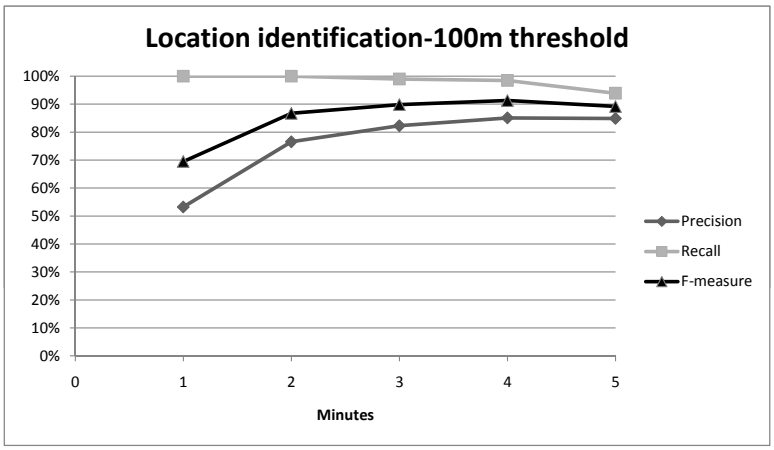

Figure 9. Location identification performance by time threshold.

To evaluate the relative performance of the online algorithm introduced here versus post processing methods, the algorithm introduced in Auld *et al.* is used as a point of comparison (Auld et al., 2009). The algorithm introduced in that work used a clustering approach of cleaned points to identify significant locations and separate them from travel during post-processing. Also rather than using a fixed distance and time threshold a more sophisticated approach was used that varied the distance threshold based on average census block size and the time threshold by speed of travel. That algorithm was evaluated on this same dataset and identified 220 activities as reported in that work. Of these 5 activities were missed (recall 97.8%) and 28 stops were identified where no activity occurred (precision 87.3%), resulting in a F-measure of 92.2%.

The algorithm introduced above was then applied and the best results were obtained using a fixed distance threshold of 100m and time threshold of 4 minutes. The algorithm identified

228 activities missing just 3 actual activity locations that were added by the participants for a recall of 98.5%. It also identified 34 activities that were removed by participants, thus not actual activities for a precision of 85.1%, resulting in a F-measure of 91.3%. While the algorithm that used post processing and a more sophisticated idea of thresholding held a slight advantage in precision, the online algorithm introduced here actually performed slightly better at recalling all significant locations. While the post processing measure did result in a slightly higher F-measure the difference was not statistically significant. It should be noted that the locations identified in the full UTRACS survey are based on the post processing algorithm due to the perceived less effort in adding 2 additional activities versus removing 6 more incorrectly identified activities.

Using a combination of automatically identifying significant locations and revisited locations was shown to provide significant time savings within the UTRACS survey. Specifically durring the survey, detected locations and matched to previously visited locations and values learned from previously visits to the location were used to default answers within the survey. Thus, from the user experience rather than entering all the information again they were given the opportunity to confirm the predicted information or change it. Figure 10 depicts the reduction in response time to answer the same questions per activity event over the 14 day period of the survey. As the results show over even a 30 day period 30 seconds could be reduced from the active entry time spent survey participants. As this figure also shows that while the answer time decreased the amount of activities they recorded remained consistent. Figure 11 demonstrates applying this same approach to trip segments based on matching to typical trips adjacent to the matched locations also significantly reduced answer time. Similar to the reduction in entry

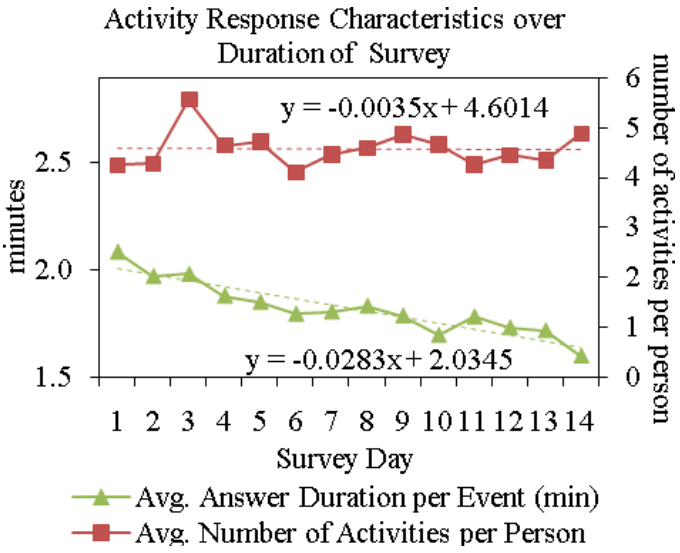

Figure 10. Reduction in answer time per activity event.

time for the activity over the entire couse of the 14 day period, the reduction in trip event entry also followed this same pattern. While the reduction was not as quick as the activity patterns, the results indicate that this will continue to reduce steadily for some period of time. Also as these results show that while the amount of trip event response time reduced, the number of trip events remained constant once again verifying that the reduction in time was not due to a reduction in activity and trip reporting. Thus by combining activity detection on the fly and inferring likely values of attributes the effort of active data entry by the user can help further reduce respondent burden.

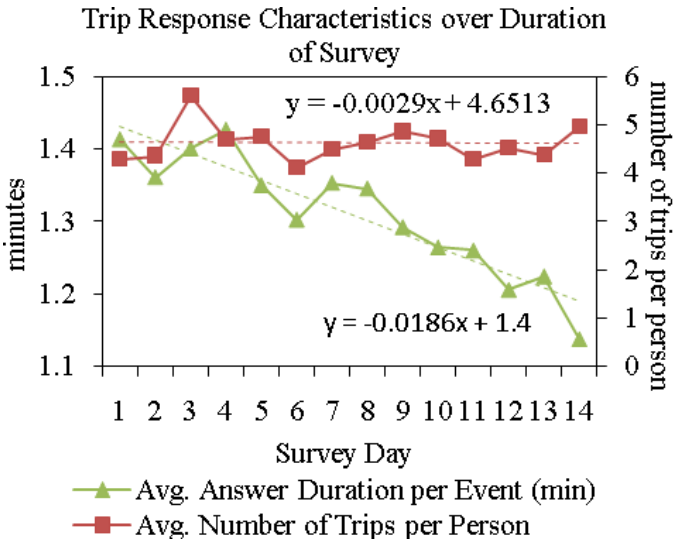

Figure 11. Reduction in answer time per trip event.

4.2.2 Discussion: reducing respondent burden through learning

With activity-based travel surveys, one of the challenges is participants are more likely to discontinue the survey if the time burden is too high on the participant. For mobile applications this is an even bigger concern since a high burden on users will likely result in the application not being used at all. Thus identifying ways to either eliminate or reduce the burden on the user while still collecting the most beneficial information is key. To address this in the UTRACS survey, one of the ways this was accomplished was to ask more questions up front in order to reduce the need for information on routine or recurrent behavior. This information combined with frequently seen patterns could then be used to auto-populate likely answers and avoid burdening the user for routine activities. A similar approach would make sense for

the solution proposed by this work as well. As demonstrated by the UTRACS survey, this approach combined with familiarity of the questions reduced the time spent by respondents on the questions by 19% (Frignani et al., 2010a).

The basic approach to reducing the effort would be driven based on a confidence threshold applied to any particular question. For each potential question, based on the other information known about the activity context and the past activity contexts, either derived through GPS processing or known with high confidence from past observations; the most likely response and a confidence in the response would be determined. If the confidence was sufficiently high the question would not be asked at all, otherwise the question would be displayed and defaulted to the expected response for confirmation. Using this approach a routine pattern like going out to a frequented place for lunch from work may not prompt any questions while visiting a new location around that same time might prompt for confirmation that this new location is a place for lunch.

While many attributes can be populated with reasonable confidence based on patterns commonly observed across all travelers, one of the advantages of long term collection is the opportunity to tailor the model to the individual. Thanks to a lengthier observation period, as the collection progresses a model can be built that better reflects the patterns of that individual. Developing a more specialized model is likely to result in higher confidence in predictions and better coverage of the number of fields that can be auto-populated. With more lengthy use of the application, a more reliable model of the individual would allow significant reduction in respondent burden.

4.2.3 Missing values

Since the goal is to collect as much useful information as possible to provide a context to mobile applications without being overbearing, flexibility is needed in collecting the data. The expectation would be that while some data about the activity may be entered/confirmed, many of the questions would not be answered. The advantage of being able to passively identify trip information and stop information is a history of this data can be collected without any burden on the traveler. Thus the user can answer as many or as few of the questions as they wish while a history of patterns continues to be collected. The result is a stream of the traveler's activity contexts with many values likely missing from when questions were unanswered. While this stream will have a rich history of the passively collected attributes the user entered data will likely be more sporadic. As a result of this formulation, being able to predict patterns despite having large amounts of values missing becomes key and remains a focus throughout the remainder of this work.

4.3 Filling in the gaps in traveler logs

This section is based on previously published work (Williams et al., 2009).

In the past, data collected in travel histories relied extensively on manual data entry by the participant. With the increase in prevalence of hand held computers and GPS devices, the detail and types of information that can be captured for travel histories are becoming far richer than ever before. While much of this data about location can now be captured passively, there is still a significant amount of detail about the planning process and personal flexibility that still must be entered manually. While these data sources share many characteristics with other

studied activity logs, the additional information on decision processes and opportunities for further enhancement with GIS offer some unique challenges as well as opportunities in behavior prediction.

The downside of collecting such a feature rich behavioral profile is that current survey techniques require participants to manually provide all planning and flexibility details, creating a significant burden on survey participants. While it would be ideal if all traveler context information were collected for the traveler being modeled for every activity and trip, the effort and burden associated with this would make this impractical for any sort of long-term collection effort. As a result the approach of many efforts has been to focus strictly on data that can be passively collected such as GPS logs (Gogate et al., 2005; Liao et al., 2007). While collecting all information manually at every step would be overly burdensome, there is likely some level of information or effort that is acceptable beyond strictly passive collection. Two possible alternatives are either reducing the data that is collected, or significantly reducing the data entry effort. In this section, we introduce a technique for improving the quality of information that can be inferred in a partially labeled sequence. Applications of the techniques introduced here would either allow less data entry or reduce the effort required for the data entry process.

Research has shown associative and sequential mining can be powerful tools for predicting future behavior and understanding past behavior. In web data mining, these techniques have been used successfully for extracting behavioral patterns for uses like personalization, pre-fetching, design analysis, and relationship mining (Facca and Lanzi, 2005). However, all of these applications focus on using *captured* behavior for predicting future actions or trends,

making little use of the captured information to infer any additional detail about past activity. Other recent work, with a more abstract link to behavior, in text mining has shown sequential analysis to be very useful in inferring labels given words before and after a gap of interest (Hotho et al., 2005). In this section we introduce an innovative application of sequential mining for reliably *inferring* past behavior for the purpose of reducing participant burden while still being able to gather a detailed behavioral history. Below behavioral inference is examined within the context of a recent travel behavior survey to illustrate the benefit this type of inference could provide even if only a portion of the data were collected.

In this section, we focus on an innovative application of sequential mining and analysis for reducing the manual effort required to record detailed behavioral histories. We examine the challenge of traveler prediction not as activity or location prediction, but as traveler context prediction. While we are also interested in projecting the context of future behavior, in this section we focus on reliably inferring past traveler context within a stream of both labeled and partially labeled travel history. We propose a novel approach to inferring individual traveler behavior and associated context from a stream of their prior activity context and the histories of others. We empirically examine how well the known context of a prior survey can be inferred when data is removed, for the purpose of developing future travel history collection techniques that require less effort by participants while still obtaining a reliable history of their behavioral context through inference.

4.3.1 Background and motivation

While there has been considerable research by urban planners on modeling traveler behavior from activity surveys for micro-simulations, it is important to recognize the differences between these two tasks. In most micro-simulations, the goal is not to simulate a specific individual, but rather to replicate behavior commonly observed by surveyed people similar to the simulated traveler. For survey learning, on the other hand, rather than focusing on prediction of common activity sequences based on the history of many travelers, the paradigm is shifted to predicting the travel context of an individual based on their own travel history and the history of others. This subtle shift facilitates moving travel prediction from the realm of urban planning to instead enabling intelligent travel surveys.

The basic concept behind activity-based analysis is that traveler behavior can be broken down into activities to accomplish a traveler's needs and travel associated with getting to the activity. Research has shown the planning of these activities happens in a fluid manner based on both personal flexibility and activity flexibility, making these aspects critical for planning decisions. As a result, the model the goal of traveler prediction is not activity or location prediction, but context prediction. Thus, the problem becomes: given a sequence of the traveler's prior context how well can the context of their next activity be predicted. Alternatively, as examined in this section, given a partial sequence of the traveler's context what can be inferred about the missing context. In the proposed model, each step in the history of the traveler form a sequence of sets describing the traveler's context progressing through time, or an enhanced activity sequence.

In web and e-commerce applications, sequential rules have been used in personalization for identifying pages or products of interest based on a user's history to reduce navigation requirements. We propose applying similar techniques to surveys so that the historical context of a traveler is used to predict their likely next step or inferring behavior based on their surrounding history to help reduce participant burden in the form of data entry requirements. In the past the amount of data entry that could actually be reduced might have been questionable as details would still have to be entered about each trip even if full details were not required; but with GPS enabled surveys becoming more widespread, a significant amount of detail about the trip can be captured passively make the task much more tractable. Recent work in transportation analysis has shown activity and travel segments can be reliably separated from GPS traces, simplifying the problem to completing a partially labeled activity sequence.

Depending on the goals and participant willingness, there are two different ways predictive models could be applied to reduce data entry requirements: auto population or selective querying. While traditionally analysis of data collected during the course of a survey is primarily done at the conclusion of the study; by collecting and analyzing the data on a periodic basis, the interim data may provide insights that can be used to improve the history collection process. Below we discuss techniques for utilizing patterns extracted from these enhanced activity sequences for reducing the burden of participants.

For auto population, the predictive model would be applied and questions about activity or travel could be pre-populated based on the user's prior history to be confirmed or changed by the participant. Consider a scenario where a GPS unit recorded a traveler making a five-minute stop

on the way to the train station. If the participant's history showed they occasionally stopped in this location for coffee, this information could be used to auto populate the purpose of the stop, their end time flexibility, and the likely planning horizon for the trip without the participant needing to enter it by hand. For longer-term surveys, this type of predictive model could be incorporated to reduce the number of questions asked. Two possible approaches would be high confidence elimination or key event querying. The principle behind high confidence elimination is to eliminate any question where the confidence that the answer is known is over a certain threshold. An alternative to this more suited for longer term history collection would be to only ask about activities or travel that are unusual compared to known patterns. In both of these approaches, while the participant still has a significant burden early on, as the data collection progresses their burden is reduced as the application learns their behavior. While learning patterns specific to a participant are valuable, due to the amount of time necessary to observe these trends, augmenting the data with the patterns of others can likely help to reduce the initial learning time.

These learning models can therefore be used to either assist or completely replace the data entry requirements of the respondent. Depending on the length of the survey and the types of attributes required, this can help to significantly reduce the respondent burden, although as mentioned the burden during the initial phase of the data collection could still be somewhat large as the algorithms learn the user's likely activity-travel patterns. However, this could further be reduced with a well-designed up-front survey of the person, which in addition to capturing socio-demographic information could also be used to identify common locations visited and

routines within the respondent's usual activity-travel pattern. The use of initial inputs of this type would likely reduce the time needed for the algorithms to develop a useful predictive model.

Value of intelligent travel history collection

In order to further reduce the burden placed on survey respondents to enable longer duration surveys, the frequency and type of questions asked of participants needs to be significantly reduced. Some routines have been developed to accomplish this, as discussed previously, by automatically detecting some attribute, which would negate the need for questioning the individual. Examples of these routines include automated trip purpose detection (Wolf, 2000) and mode identification (Tsui and Shalaby, 2006). However, these procedures are less applicable for most other attributes, which are required from the current survey. It is not obvious, for example, how planning horizons, decision variables and other attributes such as involved persons could be derived from the GPS/GIS data alone. Therefore, a learning approach is needed, which utilizes information already collected in the survey to develop patterns, which can predict the various activity-travels attributes. This section discusses some background in machine learning and some ways in which it has been applied in travel pattern prediction as well as propositions for using it to help reduce survey respondent burden.

4.3.2 Reducing the cost of missing data: Attribute Constrained Rules

Sequential pattern and rule mining have been the focus of much research, however predicting missing sets of elements within a sequence remains a challenge. Recent work in survey design suggests that if these missing elements can be inferred with a higher degree of certainty, it could greatly reduce the time burden on survey participants. To address this problem and the

more general problem of missing sensor data, we introduce a new form of constrained sequential rules that use attribute presence to better capture rule confidence in sequences with missing data than previous constraint based techniques. Specifically we examine the problem of given a partially labeled sequence of sets, how well can the missing attributes be inferred. Our study shows this technique significantly improves prediction robustness when even large amounts of data are missing compared to traditional techniques.

Frequent pattern mining of sequences has been a prominent research theme since its introduction by Agrawal and Srikant (Agrawal and Srikant, 1995), yet how to effectively use pattern-based mining for classification and prediction remains a challenge (Han et al., 2007). The problem examined in this section is given a sequence of sets of attribute values where one to all fields within a set can be missing, populate the missing attribute values in the sequence. We refer to this problem as *partially labeled sequence completion.* Two common versions of this problem are:

1. given the prior sequence, complete the missing elements of the current/next set; and
2. given a target set anywhere in a sequence complete the missing attribute values using the sets both before and after the target set

Many studies have focused on the next step prediction form, particularly for web applications such as pre-fetching and personalization (Yang et al., 2001; Mobasher et al., 2002). In this study, we examine the more general form of the problem, since it also addresses the growing number of applications that would benefit from inferring additional information about an event given both the events before and after the event of interest. An example of this would be a group of

mobile sensors that are periodically collected, where any number of the readings may be missing from any particular time step (North et al., 2008). Other work in survey design suggests that if missing elements within a sequence can be inferred with a higher degree of certainty, it would greatly reduce the time burden on survey participants (Marca et al., 2002; Auld et al., 2009).

This section presents a new form of constrained sequential rules to address the more general form of the problem, which can also be applied for next step set prediction. Attribute constrained rules (ACR) are based on traditional sequential rules that can be derived from frequent sequential pattern mining, however extensions are made to better address attribute labeled sequences with missing value data. This problem of partially labeled sequence completion is formally stated below, followed by an overview of related work and an illustrative example of the value of ACR rules and how they are mined. Our study then shows this technique significantly improves prediction robustness for even large numbers of missing values compared to traditional sequential rules using a publicly available travel survey data set.

4.3.2.1 Partially Labeled Sequence Completion

Algorithms for rule mining of sequential patterns have been a major source of interest since they were first introduced in (Agrawal et al., 1993). Much of prior work has focused on mining sequential rules for predicting future patterns, however being able to infer missing information within an observed sequence of sets of attribute values also has many useful applications that have largely been overlooked. One example of this is multi-day travel survey design, where being able to reliably infer missing values from the surrounding data set would allow respondent burden to be greatly reduced if only a portion of the data points needed to be collected regularly

rather than the full set of questions (Auld et al., 2009). To address problems such as these, we examine the more general problem of missing information within a sequence, however the technique also applies to traditional prediction as well. Specifically the problem we address is given a sequence where there are a known set of attributes that describe an event within the sequence, infer any missing values of the attributes for a target set. In this section, we formalize this constrained sequence problem and introduce a technique for mining and applying rules specifically for this task.

Problem Statement

For the problem of partially labeled sequence completion, let

$$H = \{H_1, H_2, \cdots, H_n\}$$

be a database of sequences, and let:

$$H_i =< S_1, S_2, \cdots, S_n >$$

be the sequence of sets of observations in a sequence i; where each observation set S_j is composed of 1 to m attributes $\{a_{j,1}, a_{j,2}, \cdots, a_{j,m}\}$. Each attribute $a_{j,k}$ has a discrete set of values for the k^{th} position that is shared across all observation sets S. Intuitively the sequence H_i can be thought of as a series of recordings by a survey instrument or sensor with a fixed set of discrete measures (the attributes), where at each event j all measurements are relevant, but only a portion of these measures may actually be recorded in the set S_j. Given a sequence H_{target} of

length l and a target set S_t to be completed where $1 \leq t \leq l$ and between 1 to m arbitrary attributes are missing values. Determine the values of all missing attributes $\{a_{t,1}, a_{t,2}, \cdots, a_{t,m}\}$ in S_t. Thus our goal is to use the surrounding sequence information in H_{target} to populate any missing values to complete the set S_t.

4.3.3 Related Work

With associative mining the aspect of incomplete information is of particular concern since its effectiveness degrades rapidly using traditional associative mining when there is incomplete or missing data. Addressing this problem has been a topic that has gained much attention. Ragel and Crmilleux's approach to this problem, known as missing value completion (MVC), was to create multiple views of the data so that when mining any attribute only the records that contained the attribute were included (Ragel and Crmilleux, 1998; Ragel and Crmilleux, 1999). The result was relationships could be effectively mined by only considering the relationships when a value did appear. One of the problems of this approach is that while it is very effective for associative set mining it does not work well for sequential associative mining. If this approach was applied to sequences, the only sequences that would be considered would be ones in which the attribute value of interest appeared in all sets of the sequence greatly reducing the likelihood of being able to identify more subtle patterns. Shen *et al.* examined a different approach to this problem based on using association rules and combining these with sub-frequent item sets to populate missing values (Shen et al., 2007). More recent work by Bashir *et al.* takes an iterative approach using a combination of association rules and k-Nearest Neighbor (kNN) to impute the missing values and improve the performance of the association rules in missing value

scenarios (Bashir et al., 2009). The general approach of their method is alternate between filling missing values with association rules when possible and then when no rules are applicable, kNN is used to impute remaining values. While many techniques have been introduced to handle the problem of missing data, most of these techniques are not suitable for sequential data.

One of the common aspects of the majority of association mining algorithms is they do not take advantage of variable information or variable presence which becomes particularly important in sequences with missing values. For these types of sequences, which we refer to as partially labeled sequences, if we consider the sets being observed as a set of possible attribute assignments from a portion of the set of attributes (such as instrument output) we are observing, the problem of predicting future sets can become far more well defined in terms of the possible values given a set of possible attributes with either known or inferred constraints.

While techniques have been introduced for mining sequential patterns given regular expression constraints (Garofalakis et al., 1999; Garofalakis et al., 2002), the expression constraints in these works are best suited for matching a value pattern. For example, while an expression can be defined to match any sequence of values that can be described by a regular expression, the language does not provide for a more sophisticated notion of attribute value restrictions. While some aspects of this type of functionality could be encoded to restrict attribute values to a particular value such as the regular expression constraint: $\{\text{“}a_1 = 1\text{”},(\text{“}a_2 = 2\text{”}|\text{“}a_2 = 3\text{”})\}$, this type of encoding is insufficient for constraining the general presence of an attribute if all values are not known ahead of time.

Other work such as (Pei et al., 2002) has sought to identify the specific types of constraints needed for sequential pattern mining beyond those that can be expressed with regular expressions. Their work introduces the concept of *item constraints*, which are intended to allow the presence (or absence) of a particular individual or group of items to be specified in the mining process. Given a particular query described in terms of constraints the algorithm they introduced, PrefixGrowth, finds patterns matching the constraints through a prefix extension method. While this algorithm is effective for finding patterns matching a particular query, it does not address being able to identify the set of possible constraint based rules for completing the values of a pattern in general.

4.3.4 Attribute Constrained Rule Mining

To address rules of this form we extend an idea from the text mining community called *label sequential rules* (LSR) (Liu et al., 2005; Liu, 2007). Originally introduced for analyzing opinions in text reviews, this rule form was proposed for identifying common sentence forms or templates where a type of word of interest, termed a label, would likely appear. These rules form a more constrained matching pattern through wild cards producing rules of the form:

$$\langle\{1\}\{3,*,*\}\{6\}\rangle \rightarrow \langle\{1\}\{3,4,7\}\{6\}\rangle$$

where confidence of the rule would be defined with respect to the likelihood of the right hand side (RHS) given all sequences that contain the wild card constrained pattern. Thus, if we are only interested in rules that address completing two items in the middle set, these constraints

TABLE II

EXAMPLE SEQUENCE DATABASE

H_1	$<\{a_1\}\{a_2, b_2\}\{b_1\}>$
H_2	$<\{a_1\}\{a_2, b_2\}\{a_2, b_1\}>$
H_3	$<\{a_1\}\{b_2, c_2\}\{a_2\}\{b_1\}>$
H_4	$<\{a_1\}\{a_2, c_1\}\{b_1\}>$

allow a more meaningful measure of rule confidence since the likelihood is only measured in relation to patterns that match the LSR template.

For the task of partially labeled sequence completion, we propose a similar idea for identifying templates of sequential patterns of attribute values which we refer to as *attribute constrained rules* (ACR). Whereas with LSR the confidence of rules specify how likely a generalization is about elements within a pattern, with ACR the rule's confidence specifies the likelihood of a specific attribute value or combination of attribute values given a surrounding pattern.

4.3.4.1 Illustrative Example.

In this section, we provide a set of illustrative examples of the benefit of constrained rules such as ACR and LSR. Throughout these examples refer to Table II as the sequence database.

Below we use the standard definitions of *support* and *confidence* defined as:

Definition 7 *The* ***support*** *of the sequential rule* $X \rightarrow Y$ *is the fraction of sequences in the database that contain* Y.

Definition 8 *The* ***confidence*** *of a sequential rule* $X \rightarrow Y$ *is the fraction of sequences in the database that contain* X *that also contain* Y.

For an example of how constrained rules can better represent the applicable confidence, consider the following scenario: $H_{target} = \langle\{a_1, b_1\}\{a_2,\ ?\}\{b_1\}\rangle$, where $S_2 = \{a_2,\ \mathbf{b{:}?}\}$ is the target set. The following traditional sequence associative rule would be applicable:

$$\langle\{a_1\}\{a_2\}\{b_1\}\rangle \rightarrow \langle\{a_1\}\{a_2, b_2\}\{b_1\}\rangle$$
$$[\text{sup} = 2/4, \text{conf} = 2/4]$$

Which can be interpreted as S_2 can be completed $\{a_2, b_2\}$ with a confidence of $2/4$. By contrast a label constrained version of the same rule:

$$\langle\{a_1\}\{a_2, *\}\{b_1\}\rangle \rightarrow \langle\{a_1\}\{a_2, \boldsymbol{b_2}\}\{b_1\}\rangle$$
$$[\text{sup} = 2/4, \mathbf{conf = 2/3}]$$

Where the notation $\{a_2, *\}$ indicates a set containing value a_2 for attribute a and a second value within the same set. As this example shows, by further constraining the attribute and location of pattern extension with LSR constraints, the confidence of the pattern is raised to $2/3$ or roughly 67%. With ACR this idea is extended to constrain pattern matches to particular attribute values of interest. In our example, since we are specifically interested in the value of attribute b, the ACR version of the same rule would be:

$$\langle\{a_1\}\{a_2, \mathbf{b{:}*}\}\{b_1\}\rangle \rightarrow \langle\{a_1\}\{a_2, \boldsymbol{b_2}\}\{b_1\}\rangle$$
$$[\text{sup} = 2/4, \mathbf{conf = 2/2}]$$

which is able to further clarify the confidence in populating attribute b, since it is able to discount sequence H_4 as it does not match the attribute constrained pattern. This advantage in accurately evaluating the value of the constrained sequence rule is the reason we examine ACR for partially labeled sequence completion.

The left hand side of Figure 12 shows the frequent sequence graph using a minimum support of 40% for Table II along with the support counts for each frequent sequence. All frequent ACR antecedents can be easily identified from the frequent item sets by expanding the remaining possibilities. For example the following antecedents can be generated from the frequent item set$\langle\{a_1\}\{b_2\}\{b_1\}\rangle$:

$$\langle\{a_1\}\{b_2\}\{b_1\}\rangle \Rightarrow \left[\begin{array}{l} \langle\{a:*\}\{b_2\}\{b_1\}\rangle, \langle\{a:*\}\{b_2\}\{b_1:*\}\rangle, \\ \langle\{a_1\}\{b:*\}\{b_1\}\rangle, \langle\{a_1\}\{b:*\}\{b:*\}\rangle, \\ \langle\{a_1\}\{b_2\}\{b:*\}\rangle, \langle\{a:*\}\{b:*\}\{b_1\}\rangle, \\ \langle\{a:*\}\{b:*\}\{b:*\}\rangle \end{array}\right.$$

As this example shows, due to the combinatorial growth of the attribute constraint sets this problem quickly becomes impractical for datasets with a large number of attribute values or lengthy sequences if all completion rules are considered. For example with even this small example, the 17 frequent sequences have over 40 potential ACR antecedents. For the problem as stated in Section 4.3.2.1 there are some properties that can be taken advantage of to reduce the number of ACR antecedents. Specifically, one of the key features of the problem we address is that every observation set S_j is composed of the same m attributes $\{a_{j,1}, a_{j,2}, \cdots, a_{j,m}\}$, and

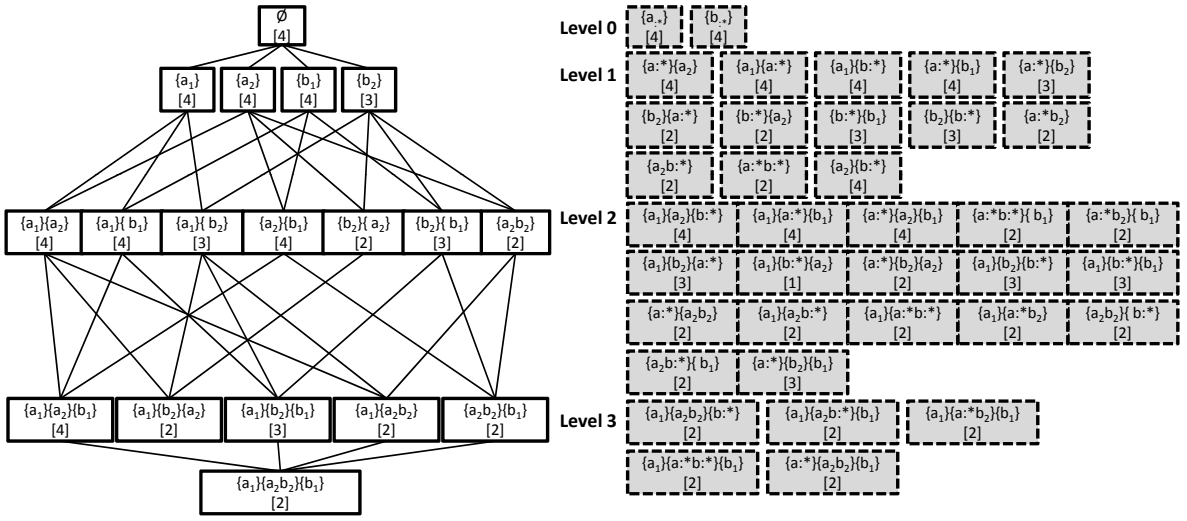

Figure 12. ACR Frequent sequence graph.

only one target set is examined at a time. The implication of this fact is the property that for any set S_i, there is a value (whether known or not) for every attribute $a_{j,i}$. This property means that while the number of possible antecedents may grow quickly, the only ones that need to be kept are those with all constraints within a single set within the sequence. Once all possible ACR antecedents for the frequent pattern sets have been enumerated, the support for all the patterns can be updated with a single pass of the data set. By adding this subset and associated support to the frequent sequence graph as shown in the right hand side of Figure 12, ACR predictions and completion can be quickly determined using just the information in the extended frequent sequence graph. Note that while not shown in the figure, links would also exist between the ACR antecedents and the frequent sequence completions in the graph.

4.3.5 Experiments

4.3.5.1 Experimental Setup

Data

To evaluate the proposed ACR technique, we chose the 2001 Atlanta Household Travel Survey for several reasons. This dataset contains a large number of sequences of sets that are known to have a strong relationship between the entire set of attributes at each step and their ordering, making it well suited for sequential set prediction. Second, the type of data collected in this survey is very similar to one of the proposed applications of this type of partially labeled sequence learning, reducing survey participant burden (Marca et al., 2002; Auld et al., 2009). Demonstrating that a significant portion of this type of data can be removed (i.e. a number of survey questions reduced) while limiting the impact on predictions is a significant step in showing the feasibility of this type of application. Finally, this data set represents one of the larger publicly available data sets of this type, making the results of this study open to competitive comparisons by other work in this area.

The 2001 Atlanta Household Travel Survey was conducted from April 2001 through April 2002 on behalf of the Atlanta Regional Commission (ARC) (NuStats, 2003). The data consists of travel activity information for 21,323 persons from 8,069 households and 126,127 places visited during the 48-hour travel period. This survey focused on observed activity type, timing, and travel associated with each person's activity schedule captured over a 2 day period. The survey captured a wide range of in-home and out-of-home activity types which were broken down by a high-level classification. The survey captures over 250 attributes relating to the

travel, activity, and demographic characteristics of the individual for each activity sequence that was recorded. The data is structured such that each event corresponds to an activity in the person's schedule with the set of attributes corresponding to the characteristics of the activity and travel associated with getting to the activity.

In the experiments below, we focus on a subset of 6 of these attributes: activity type, mode of transportation, arrival time, departure time, activity duration, and traveler age. These attributes were selected as they represent a mix of information about the type of activity, the travel, relative time the activity took place, activity duration, and features of the person involved that have been shown to be highly related both intra-event and inter-event in predicting traveler activity patterns (Timmermans, 2005; Ettema et al., 2007). Thus, the dataset can be thought of as a database of sequences of events with the set of attribute values at each event being highly related. For the subset of the data set we use, there are 49,695 sets of activity information, with an average sequence length of just over 7.4 sets. Additional information about the data set can be found in Appendix A.2.

Methods

In the results below, we present the results of both the ACR technique, introduced in this section, and traditional sequence rules for a comparative baseline. As both rule-based techniques are frequent pattern based, which is deterministic for a given minimum support threshold, in all experiments below the same set of frequent patterns were used for both the traditional sequential mining and the ACR mining to ensure a fair comparison. In all experiments, both

sets of rules were mined using the same minimum confidence threshold, and only rules that produced at least one target item in the target pattern were considered.

To generate predictions given a set of many potentially applicable rules, a ranking prediction scheme was utilized. Specifically, the rules were ranked in order by confidence, number of target productions, support, antecedent length, and finally a string based comparison to ensure repeatability if all other factors were equal. The productions of the top ranked rule were then applied to the target set, the remaining matching rules were then re-ranked as the number of target productions may have dropped due to the previous rule's productions. The rule with the highest rank of those remaining was then applied and this process continued until either a prediction had been made for all target items or no rules remained.

As described in Section 4.3.2.1, the problem of partially labeled set completion involves taking a sequence and trying to fill in or predict items within a single target set within a sequence. Since the problem of partially labeled set completion can take the form of predicting anywhere from a single item in a target set to all items in the target set, the results below reflect the average of all possible combinations of the target pattern in all possible positions for the target set. Where target pattern means: the set of attribute values in the target set that are being evaluated. Thus in the experiments below, for the target set any attribute value that is not specifically of interest as specified by the target pattern retains its original value

for determining matching rules. For example if the target pattern included attributes a and c ($S_T = \{a_T c_T\}$). In testing the sequence:

$$\langle\{a_1 b_2 c_1\}\{a_2 b_2 c_2\}\{a_1 b_1 c_2\}\rangle$$

If the target set was S_2 for the sequence, the test sequence would thus be:

$$H_{target} = \langle\{a_1 b_2 c_1\}\{a_T b_2 c_T\}\{a_1 b_1 c_2\}\rangle$$

In the base experimental data set described above, no attribute values were missing. The missing data scenarios were created by randomly removing the specified percentage of values from both the training and test sets for any attribute appearing in the target pattern. All experiments below were run using a minimum support threshold of 60% for frequent patterns and a minimum rule confidence threshold of 80%. To ensure the significance of our results, all results shown are the average of a 10 times cross-folding methodology.

4.3.5.2 Experimental Evaluation

ACR vs. Sequential Rule Comparison

In the first set of experiments, we examine the impact of missing data on each rule mining technique. Figure 13 portrays the recall of the missing attribute values. In the figure, *ACR-Avg* and *SEQ-Avg* represent the ACR results and the traditional sequential rules results respectively averaged over all possible number of target items in the target pattern. As these results show,

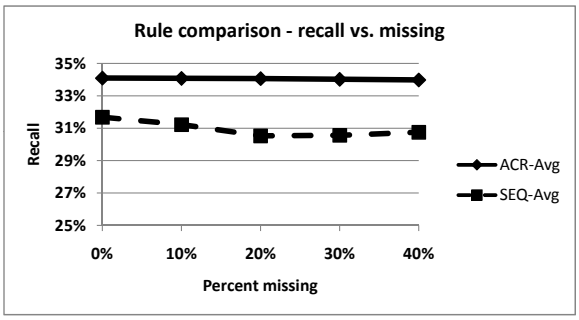

Figure 13. Comparison of recall for ACR and traditional sequential rules as the percent of missing data increases.

the ACR technique produces both a higher overall recall as well as less degradation as values are removed and the data becomes sparser.

Since higher recall can indicate an increase in the quantity of predictions at the cost of accuracy, it is important to consider the precision as well. As Figure 14 shows, the boost in recall from ACR comes at a trade-off of less than a .3% drop in precision. While this small drop is statistically significant, in practice the benefit of the additional 3-3.5% of attribute values with good predictions (recall) is likely to outweigh this small drop in precision for most applications.

The combined performance, as evaluated using the F-measure, is shown in Figure 15. As these results demonstrate, the ACR technique results in far better combined predictive robustness compared to traditional rules as the amount of missing data increases.

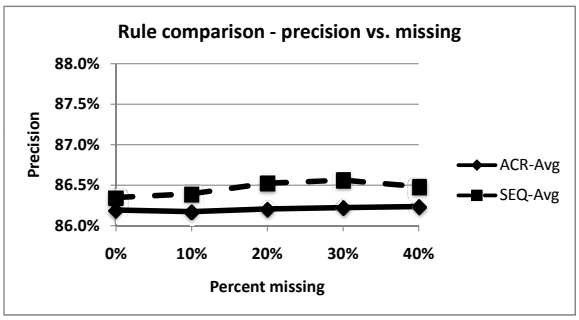

Figure 14. Comparison of precision for ACR and traditional sequential rules as the percent of missing data increases.

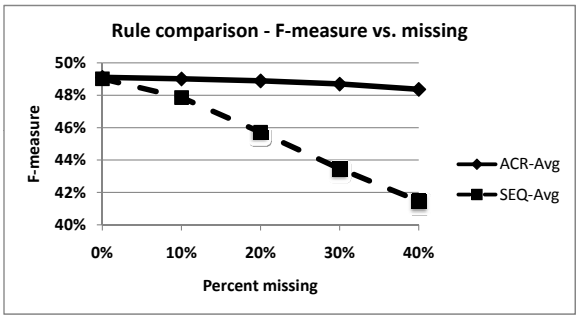

Figure 15. Comparison of F-measure for ACR and traditional sequential rules as the percent of missing data increases.

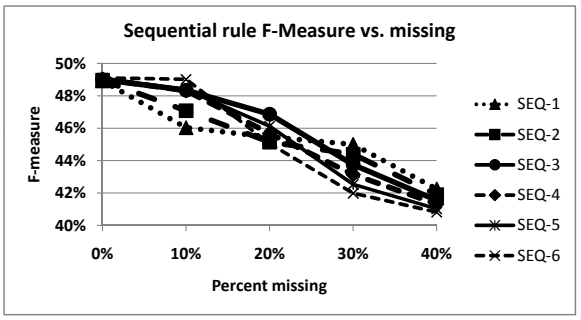

Figure 16. Comparison of differences in predictive performance for traditional sequential rules for different target set sizes as the percent of missing data increases.

Number of Target Items Comparison

In the next set of experiments, a more in depth look is taken on the impact of the number of items trying to be predicted in the target pattern. In these results, *ACR-**X*** and *SEQ-**X*** represent the ACR results and the traditional sequential rules results respectively averaged over all possible target patterns with ***X*** number of target items in the target pattern. Thus, *ACR-3* would indicate the average performance of the ACR technique averaged across all possible target patterns with exactly 3 target items. The reason this is of interest, as Figure 16 shows, is that the number of items that are trying to be predicted can have a significant impact on how missing data affects prediction performance. As the results demonstrate for traditional sequential rules, while a small amount of missing data (10%) has a greater impact when predicting fewer items; as the amount of missing data increases (30%) this relationship is reversed.

As Figure 17 shows, with traditional sequential rules, as the number of values needing to be predicted increases, it becomes increasingly harder to recall all of the target values and becomes

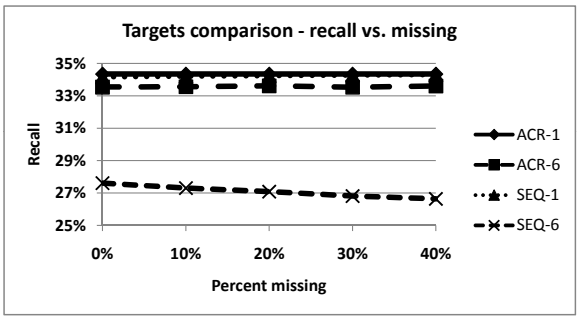

Figure 17. Comparison of recall for ACR and traditional sequential rules for different target set sizes as the percent of missing data increases.

even harder as the amount of missing data increases. Comparing this to the rules produced by the ACR technique shows that while this is still the case; the attribute constraints significantly increase the recall for this scenario even without missing values. Intuitively this is because in a less restricted target set, the attribute constraints better capture the likelihood of the predicted attributes all occurring simultaneously in the sequence than the unconstrained form of the rule.

A look at the precision for this same scenario, Figure 18, shows perhaps an intuitively unexpected result that the precision when predicting a full set is actually slightly higher than when predicting a single target item and furthermore increases slightly with some missing data. The reason for the higher precision with more target items is due largely to a smaller percentage of attribute values actually being predicted (as reflected in the recall) and in this case, is likely in part due to a feature of the data set such that some attribute values are easier to predict than others (not shown in this work). Likewise the small elevation in precision associated with a percentage of the elements being removed likely reflects the benefit of randomly reducing the

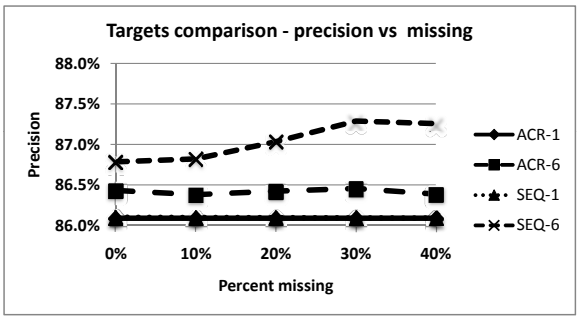

Figure 18. Comparison of precision for ACR and traditional sequential rules for different target set sizes as the percent of missing data increases.

appearance of some of the less frequent items which may have some similarity to noise when considering full set prediction.

Finally, the F-measure comparison of combined prediction performance is shown in Figure 19. As the results show, with traditional sequential rules while a small amount of data can be removed (less than 10% for this data set) with limited impact on full set prediction; as the amount of missing data increases beyond this the performance quickly degrades. Single target prediction displays a slightly different trend being much more affected by even a slight increase in the amount of missing data, but being slightly more resilient than full set prediction as the amount of missing data increases beyond the initial amount.

This general pattern for traditional sequential rules show that the fewer the number of target items, the more significant any increase in missing data becomes; but also the less effected by subsequent increases in missing data is further illustrated in Figure 16. In contrast, the ACR technique proves much more resilient in either scenario as the amount of missing data

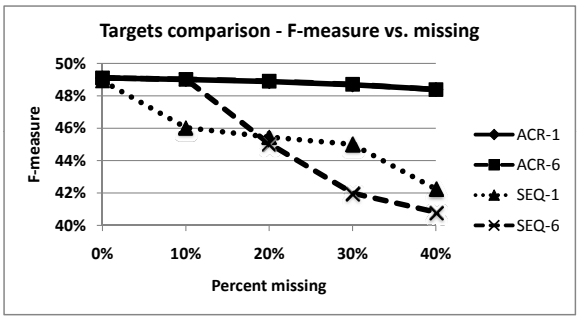

Figure 19. Comparison of F-measure for ACR and traditional sequential rules for different target set sizes as the percent of missing data increases.

increases, demonstrating nearly identical balance in predictive performance in both scenarios as the amount of missing data increases. This same nearly identical F-measure trend was observed for all target set sizes with the ACR technique (not shown).

4.3.6 Discussion

In this section we introduced attribute constrained rules, a technique for mining and better estimating rule performance for partially labeled sequence completion. As our results demonstrate, this technique shows significant promise for accurately predicting sets of attributes within a sequence with missing data compared to a traditional sequential rule-based approach. In the context of survey applications aimed at reducing the time burden on participants such as those described in (Marca et al., 2002; Auld et al., 2009); this represents a significant time savings opportunity. Specifically, the implications of the results presented in this section are that rather than needing to ask the full set of questions for each event as is currently done; a participant could be asked a much smaller portion of the questions with minimal impact on the benefits

gained by pre-populating their likely responses. In the context of other applications such as mobile sensor data, this might represent a chance to reduce costly communication without reducing the ability to reliably predict missing values.

While the technique introduced is well suited for completing multiple attributes within a set of a sequence, a heuristic take on this technique may be necessary if it were to be used for predicting multiple sets simultaneously due to the combinatorial growth of possible ACR antecedents. One area worthy of future exploration would be ways this type of approach can be adapted to handle streaming data. Finally, a study is underway to confirm the benefits of this technique in practice for interactive survey applications such as those described above.

4.4 Activity pattern transferability

One of the main reasons transportation planners have looked at modeling travel patterns based on an activity approach is its basis on behavioral patterns rather than trips alone. The input for these models are then collected through activity-based surveys. Finally, planners take the information gathered from the surveys to build activity models which are then used to project travel patterns for people in the area of study. In doing this, a key assumption is made that activity models can represent the basic patterns of people within a similar area. The validity of this approach for tour type scheduling models such as the model used in this work have been demonstrated when applied to the same metropolitan area (Doherty and Mohammadian, 2007). While this approach has been verified within the same metropolitan area at an aggregate level, the ability to use activity patterns within one city to project the activity patterns within another metropolitan area has been largely unexplored.

Recent work has examined various forms of transferability of travel data across regions. Mohammadian and Zhang examined this with respect to transferring national household travel survey data across cities (Mohammadian and Zhang, 2006; Mohammadian and Zhang, 2007). Their studies demonstrated transferring household characteristics, variables of land use, and transport network characteristics to be an effective way of estimating distributions in the absence of extensive surveys in a region.

Arentze *et al.* explored the theoretic transferability of activity patters across cities using the Albatross rule based model (Arentze et al., 2002). Their approach to evaluating activity transferability involved building a model on one city and using the model to compare the predicted travel demand from the resulting activity patterns on a second city. At an aggregate level the predicted demand was satisfactory compared to models built specifically for these cities. However, it is important to note that while aggregate travel demand was similar no comparison of the similarity of activity patterns across the cities was performed.

Other studies by Timmermans *et al.* have examined the similarity in travel patterns across different countries (Timmermans et al., 2003). An interesting aspect of their work was how different cultures, urban structures, transportation networks and accessibility differences affect activity and travel patterns. Their work found that differences in relative location and transport network were far less important than demographics, economics and social aspects. While this study demonstrated there were some similarities in aggregate travel patterns across the cities in these various countries, little analysis was made at how well a model built on one city could project patterns at another city when considering these differences.

While these studies have shown evidence of the potential to transfer activity patterns from one city to another, none of these studies have directly demonstrated that this approach is in fact valid. This section extends their ideas and addresses this question and further examines the transferability at an individual level rather than a macro level comparison of activity patterns. The contribution of this section is demonstrating that the micro patterns of one city can be used to help augment the observations within a different city for more accurate micro level predictions. While this work does not explore if this relationship holds across significant cultural differences, it does show that these patterns can be transferred across varying transportation networks and urban structures within a selection of cities within the United States and Canada.

4.4.1 Background and motivation

The principle idea behind activity modeling is that by understanding the activity priorities of travelers a model can be built to project what activities they are likely to do over a fixed period of time. Since these priorities are expected to vary only a limited amount across people of the same region, a non-individual specific model of activity execution serve as a reasonable representation of typical travel behavior. The travel of a person can then be viewed as the trips the person takes to accomplish their activities. Because the underlying activity needs are the factor driving travel rather than the location of activities and the underlying transportation network, in theory asserts similar activity patterns will be seen in people regardless of their environment.

As prior work has shown, by taking the patterns of limited group of people within a region, a general model of typical behavior can be applied to project the travel patterns across many

different locations within the same metropolitan area. One of the limitations of this work has been that without evidence that these patterns can be reliably transferred across regions, a survey of activity patterns in the region in question must take place in order for a model to be built for that region. Demonstrating that these patterns can be transferred across metropolitan areas represents a significant step forward for planners in being able to reduce the amount of data that must be collected in a region before a reasonable model of the area can be created.

At an individual level, while past studies have focused on how well these models applied at an aggregate level, an approach based on this general principle can also be applied to learn the specific activity patterns of individuals. As also shown in these sections, while the history of a single individual can be used to make predictions of that person's traveler context; a more robust model can be built faster if there is a base of activity histories of the region. Thus, the implication of this at an individual level is that unless a travel survey has been conducted in a person's region, the model is forced to rely on the individual's history alone resulting in a much larger amount of data to be collected about the individual before a meaningful model can be built. However, if activity patterns can be shown to transfer across regions the implication is that the activity patterns of the individual could be quickly learned in any metropolitan area around the country regardless of the availability of an activity survey for that area. This has significant implications for mobile platforms as it means applications could be used in a much larger portion of the country than otherwise possible. It is this desire that motivated this investigation of transferability of activity patterns.

4.4.2 Evaluation

As discussed above, the basic idea behind activity pattern transferability is that the patterns from one metropolitan area can be used to predict the patterns for another metropolitan area. To evaluate this, a series of experiments was conducted to examine multiple aspects of the transferability of activity patterns across a selection of cities. The details of the experiment design and results are given below.

The main purpose of these experiments are to compare how well survey data from other cities can be used to predict the activity behavior in a city where little if any data has been collected. The basic principle behind the experiment design was to verify the transferability of activity patterns by building a predictive model on the activity surveys from one city and use this model to predict the activity patterns of another city.

4.4.2.1 Data

To make an evaluation of how well activity patterns transfer across cities required both activity surveys from multiple cities and a common definition of activity data. To fulfill these requirements, additional data sources from existing studies were used to supplement the UTRACS data collected as part of this work. For evaluation, this thesis focused on three sources of outside data. Specifically the data from the Toronto CHASE survey, the 2001 Atlanta Household Travel Survey and the 2002 Anchorage Household Travel Survey were used. A general description of these data sets is given in Sections 3.3.3.2 and Section 3.3.3.1. These data sources were selected as they provided a good mix of metropolitan populations and differences, survey sizes, and similarity in the data collected.

One of the most significant challenges in transferring activity patterns across cities is data compatibility. While numerous activity surveys have been conducted, it is rare that any two surveys capture exactly the same information. These differences are primarily due to two factors: variations in survey goals and a lack of standardized naming even for surveys conducted by the same companies (NuStats, 2003; NuStats, 2002). As a result, to compare activity patterns between cities a common vocabulary must be established. For the experiments conducted within this section all four activity surveys were mapped conceptually to the vocabulary established in our design of the UTRACS survey. Since the primary focus of this section is related to activity patterns this section only examines the transferability of these attributes, however a more generalized for of conceptual alignment of activity surveys and being able to automate this process is discussed in Section 4.5.

For this section the following fields were compared:

- Location
- Activity
- Mode
- Arrival time
- Departure time
- Duration
- Duration flexibility
- Time flexibility

- Spatial flexibility

4.4.2.2 Methods

To evaluation the transference of activity patterns a comparison of predictive quality was used. Each of the data sets was used as a test data set for all other data sets and was tested as the training data set for all other data sets. In addition to these, each data set was tested on itself with a ten times cross-folding methodology.

The method for executing the tests was to train the classifier and then for the test data set evaluate each traveler's sequence separately. The training data was built by taking each user sequence in the training set and splitting the sequence by home-based tour. To capture a new tour's start being dependent on the activity last completed before the current home-based tour began each tour after the first would start with the last two elements of the previous sequence. Thus each training user's sequence was split such that each new sequence would begin with an out of home activity followed by a home-based activity followed by all activities up to and including the next home-based activity. This method was used to break up long sequences such as those in the multi-day CHASE and UTRACS data where an activity sequence for the entire time period could contain over 80 activity sets.

The evaluation was done so that in evaluating the test user, the activity sequence of the user was evaluated in a stepwise fashion. In other words, for the first prediction of the user there was no history to base the prediction on, but for the second prediction there would be one set of history in the sequence to base the prediction on. Similar to the training approach, the user's history used in testing was the current home-based tour and the previous home-based

tour. This was continued through the end of the sequence where each prediction of the n^{th} step of the tour was based on $n-1$ sets in the user's current tour plus the previous home-based tour. This process was repeated for each user in the test set and the average of all of these predictions is presented.

One of the challenges of the learning activity patterns of these data sets is that all four of these data sets contain numerous missing values from survey participant's not answering all questions. If the user sequences with any non-response were removed from the data set the result for most of these sources would be less than 50% of the user sequences would remain. In Section 4.3 it was demonstrated that the ACR method introduced in the section performed significantly better than standard sequential rule mining when the elements of data missing were completely random referred to as MAR. In these experiments the data sets are used as is, meaning the distribution of missing values was based on actual user non-response rather than completely at random. To further demonstrate the benefits of the ACR technique in this environment, the experiments were conducted with both standard sequential mining and ACR and the results are presented below.

4.4.2.3 Experimental evaluation

For the following experiments the data from the UTRACS (Chicago), Atlanta, Anchorage and CHASE (Toronto) activity surveys were used to evaluate how well activity patterns could transfer across regions despite the significantly different transportation networks in each of these cities. The data from all of these surveys were mapped to the common data schema and all experiments executed in that common format.

Within these experiments, the primary goal is to determine if surveys from one region can be used in a different region with little drop off in the quality of the predictions compared to a model built on the same city. The approach for verifying this transferability is similar to that used in Arentze *et al.* where the data of one city is used as training data for a second city and the results of that test being compared to the results of a city being trained on its own data (Arentze et al., 2002). For their study the evaluation was made at an aggregate level to compare distributions, by contrast here we analyze the transferability of patterns at an individual scheduling level. For these experiments the results of training and testing for the same city were performed using a ten times cross-folding methodology. The cross-folding methodology was executed such that each fold had roughly the same number of travelers although the number of activities per fold could be vastly different depending on the users selected for particular fold. For all other tests of training based on one city and testing on a second city were carried out by using the entire survey set of the training city and tested on the entire survey set of the target city.

In making the predictions, once all rules were identified the same ranking scheme discussed in Section 4.3 was applied. Predictions were then made by applying the top matching rule, updating the target set and finding the next highest rule whose consequent did not contradict the values previously completed. This was repeated until either all rules had been exhausted or all fields in the set being tested were completed. The values shown in the tables are based on selecting each algorithm's (Sequential or ACR based) parameters of support and confidence thresholds that yielded the maximum F-measure. The precision and recall metrics shown are based on this configuration for the maximum F-measure.

TABLE III

ACTIVITY PATTERN TRANSFERABILITY - F-MEASURE

SEQ

		Test			
		UTRACS	Atlanta	Anchorage	CHASE
Training	UTRACS	**0.348512**	0.298107	0.306271	0.440383
	Atlanta	0.207837	**0.294993**	0.309354	0.44916
	Anchorage	0.187883	0.290179	**0.311455**	0.423933
	CHASE	0.191884	0.289139	0.308722	**0.50395**

ACR

		Test			
		UTRACS	Atlanta	Anchorage	CHASE
Training	UTRACS	**0.461126**	0.447509	0.427923	0.575487
	Atlanta	0.333476	**0.443616**	0.441377	0.60563
	Anchorage	0.41161	0.461426	**0.546549**	0.57918
	CHASE	0.276572	0.417372	0.426867	**0.625694**

The experiments were executed so that every city was treated as a training and test set for all of the cities. Table III contains the F-measure results of these tests. Within all of the tables in this section the results of standard sequential rules (SEQ) are shown above the results of the same experiments using ACR rules. The rows represent how well the survey of a particular city was at predicting the patterns in other cities. Likewise the columns represent how well various cities performed at predicting the patterns of that particular city. Across the diagonal marked in **bold** are the results of training and testing on the same city.

TABLE IV

ACTIVITY PATTERN TRANSFERABILITY - PRECISION

SEQ

		Test			
		UTRACS	Atlanta	Anchorage	CHASE
Training	UTRACS	**0.37588**	0.366469	0.39514	0.574287
	Atlanta	0.234009	**0.366208**	0.388359	0.592757
	Anchorage	0.216313	0.357425	**0.397476**	0.536095
	CHASE	0.218641	0.352076	0.391261	**0.642939**

ACR

		Test			
		UTRACS	Atlanta	Anchorage	CHASE
Training	UTRACS	**0.537679**	0.57346	0.551117	0.803447
	Atlanta	0.35852	**0.522062**	0.493739	0.750434
	Anchorage	0.41161	0.461426	**0.546549**	0.57918
	CHASE	0.305204	0.500242	0.4859	**0.81879**

In examining the results, one of the first things that becomes apparent is how much better ACR rules are at handling missing values compared to standard sequential rules. A closer inspection shows that for every single training and test combination the ACR method outperformed standard sequential rules. These results are of particular interest as the comparative analysis shown in Section 4.3 compared the performance of these two different types of rules with missing values inserted at random. As the results here demonstrate, the same advantages shown with MAR holds true when the distribution reflects actual observed distributions which likely have some fields having far more missing values than others.

TABLE V

ACTIVITY PATTERN TRANSFERABILITY - RECALL

SEQ

		Test			
		UTRACS	Atlanta	Anchorage	CHASE
Training	UTRACS	**0.325033**	0.251241	0.250036	0.357116
	Atlanta	0.18693	**0.24698**	0.25706	0.361569
	Anchorage	0.166058	0.24423	**0.256051**	0.350583
	CHASE	0.170963	0.24529	0.254941	**0.414372**

ACR

		Test			
		UTRACS	Atlanta	Anchorage	CHASE
Training	UTRACS	**0.40414**	0.36692	0.349743	0.448294
	Atlanta	0.311702	**0.385673**	0.399057	0.50767
	Anchorage	0.41161	0.461426	**0.546549**	0.57918
	CHASE	0.252851	0.358057	0.380624	**0.506294**

As these results in Table III demonstrate, it is possible in nearly all of these cases to find a secondary city which can train activity patterns for the first nearly as well as the city itself. For Atlanta there appear to be two cities that can actually slightly improve the performance over using Atlanta alone, however the difference is not statistically significant. The one exception to this was Anchorage which there was a drop off in rule performance. One of the more interesting observations related to this is the differences between the ACR and sequential rule techniques in which cities had difficulty finding a good outside training city. While with ACR Anchorage is the only one, with standard sequential rules while Anchorage is not an issue the UTRACS and

CHASE surveys have this issue. One explanation that would explain this for the CHASE data is it contains the highest number of missing values making it more difficult for the standard sequential method to recognize the underlying pattern when applying rules. For the UTRACS data while there are not as many missing values the values are more distributed across multiple fields than the other data sets.

In the next set of experiments a comparison was made of using a model built just based on the user's individual history against a hybrid model that was seeded with general patterns observed in survey data. As this experiment is trying to demonstrate that this work could be applied in a city other than one with survey data, these experiments were conducted on the UTRACS data, but the general patterns were mined from the Anchorage data set. Figure 20 shows the comparison of the F-measure, or balanced prediction score, of these two different approaches day by day for the 14 day period of the UTRACS survey. The points charted are the average score of predicting all fields of the activity context for the day averaged across all participants for that day. The series labeled "Missing 0%" represents the model of the user with the data as entered by the survey participants. Thus there are fields that are missing due to participants not providing an answer, but no additional data has been removed. The series labeled "Missing X%" chart the performance using the same set of data but with X% of the values removed at random. The purpose of this is to project the amount of missing data that may be typical for a user's participation in an application setting rather than as a survey participant. As the chart shows, as expected the best results are obtained if the full history is used in modeling and with as short a learning time as 2 weeks a good model of the

individual can be created unlike the GPS based models alone which typically have a 4 to 6 week training period. Despite these good results, in an application setting it is unlikely a user would be willing to enter everything about their travel for a 2 week period. To approximate typical behavior series are shown for the user only answering 10% and 20% of the questions respectfully. As the results show, when only that limited amount of information is provided it is difficult to learn a model of that individual quickly. This problem referred to as the "cold start" problem occurs due to limited training examples being available until enough history is gathered. To address the cold start problem we introduce a hybrid approach to this problem that uses a combination of an individualized model seeded with patterns. The technique used for combining the two models was the individual model first attempted to predict all activity context fields, if any fields remained the general model was then applied. As the Hybrid 80% and Hybrid 90% series show, this approach greatly improved the performance despite having very few observations of the traveler. Also when this same model was applied to the 0% missing case a slight improvement was shown using this same hybrid model due to the slightly higher recall thanks to the much larger database of patterns (Not shown). One of the limitations of the 14 day survey period is it is difficult to determine the rate of gain for the amount of history gathered when so few fields are answered. This would be an area to further explore with a larger scale long term study, which was unavailable at the time of this writing.

In Table IV the precision of the predictions for both ACR and standard sequential rules can be seen. Once again ACR dominates standard sequential rules across the board. An interesting thing to note in these results is that the UTRACS survey as a training set performs

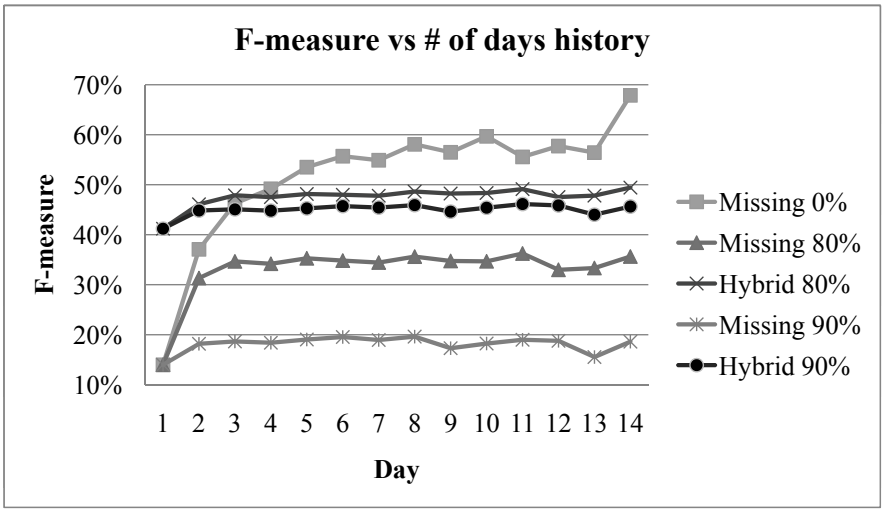

Figure 20. Comparison of F-measure for individual model vs hybrid model with missing data.

better for all test combinations than any other secondary city test set. Whereas looking at the recall in Table V, the opposite is true. For recall the UTRACS survey displays some of the worst recall. One explanation for this would be that the lengthier UTRACS survey had more examples to better capture activity sequences, however due to the survey's relatively small size its recall was not as high as other studies. This assertion about why the recall is low is further supported by the CHASE data set also suffering from lower recall and it is also one of the smaller studies. An interesting observation from these results is that the CHASE data set has the poorest transferability. This could warrant further study to better determine if this is potentially related to differences between preferences between the United States and Canada, the differences in survey participants or perhaps just an artifact of the translation to the common schema.

4.4.3 Discussion

As shown above, while generally transferring patterns from one city to another produced good results, the selection of the training data could significantly impact overall performance as well as the precision versus recall mix. Even so, the implication of this is that even if a user is in a city that other histories are not available, survey data from a second city could be used instead yielding good results despite the lack of city specific data. Furthermore, the use of lengthier surveys such as UTRACS could even produce better results than models built upon the same city. The reason for this is likely due to these extended surveys capturing more detail in the sequential dependencies of activities due to the longer observation period than the surveys based on two days alone. An implication of these findings, beyond the scope of this work, is that by using a similar approach, it may be possible to reduce the required number of survey participants while still being able to obtain a reasonable model of the travelers.

4.5 Automated ontology alignment of transportation surveys

In this section we examine the problem of ontology alignment of transportation surveys. As shown above, by aligning survey data additional data sources beyond the original can have a significant impact on improving the learning of traveler patterns. However aligning surveys completely by hand is a very tedious task. In this section we examine the use of automatic alignment techniques to reduce the number of fields that must be aligned by hand between two data sources. Travel surveys represent a particularly difficult challenge for automated alignment techniques since it involves aligning coded values and numeric range alignment that has been largely overlooked in prior ontology alignment work. The primary contribution of this section is

an alignment algorithm that combines concept and structural alignment techniques with novel numeric alignment methods that are shown to significantly outperform existing techniques for this type of problem.

4.5.1 Introduction

In this section we examine a hybrid of ontology alignment approaches. Specifically a combined approach is examined that utilizes techniques from concept-based and structure-based ontology alignment, suitable for data sources where instance data is largely encoded making instance-based techniques that work on the raw instance data impractical. Two particular problems of this type of data are examined. First, schema matching problems that involve similar concept names in multiple places within the structure particularly of encoded values, thus requiring structural context for resolution. Second, alignment of range concepts will be examined. For example, recognizing that the concepts "'1-5 xxx"' and "'6-10 xxx"' should be matched to the concept "'1-10 xxx"'.

The scenario that is examined in this project is automatic joining of metropolitan travel surveys [1]. Unlike the previous section which focused on a handful of attributes that capture activity context, here the task examined is aligning the complete transportation surveys as a way of potentially augmenting the traveler profile beyond just that collected in the activity context. A feature of the travel survey data sets is that the instance data is highly encoded to facilitate standardized answers within a survey, but the standardizations are not present across

---

[1] http://www.surveyarchive.org/

surveys. While all the cities are trying to capture roughly the same information the surveys are dissimilar enough to require a more sophisticated approach to matching concepts across structures based on linguistics and heuristics. Thus, this data set is challenging for traditional ontology alignment methods that look for distinct textual based matching.

4.5.2 Related work

Ontology alignment has been studied extensively over the past several years. Numerous techniques have been applied to address this problem focusing largely on concept and structural techniques (Rahm and Bernstein, 2001; Noy, 2004; Shvaiko and Euzenat, 2005). More recent work has examined this task in the context of data mining, where schemas are not as conceptualized and often lead to more ad hoc and often more challenging alignment problems (Hovy, 1998; Hu and Qu, 2007). These problems are particularly prevalent in the semantic web and have led to numerous similarity techniques (Fossati et al., 2006; Euzenat and Valtchev, 2004; Stoilos et al., 2005; Melnik et al., 2002; Marshall et al., 2006; Chen et al., 2006). More closely related work by Edwards *et al.* focused on learning similar topics across web pages (different structure, similar content) (Edwards et al., 2002). Berendt *et al.* examined learning frequent concept relationships across web pages (Berendt et al., 2002). Finally, in work by Tresp *et al.* they introduce the concept of probabilistic alignment rather than binary alignment (Tresp et al., 2008).

4.5.3 Survey alignment

To address the challenges described above, a combination of both structure-based and concept-based techniques is examined. In addition, new techniques are introduced to bet-

ter account for numeric range attributes. In the section below, an algorithm is introduced to address this problem. The basic idea behind our algorithm is to score the quality of a match based on three factors: parent level match, same level match, and child level match. These factors are then combined in a simple linear combination to determine the overall quality of the match, where each of the three raw scores are in the range [0,1]. The detailed description of the intuition and technique behind each of the three levels can be found below in this section; an extended discussion of the performance characteristics of this technique can be found in Section 4.5.5. Below is the pseudo code for the proposed algorithm:

```
getMatchScore(sourceE, targetE)
  // Parent level match described in Section 4.5.3.1
  parentLevelScore =
    maxOf(
      getMatchScore(sourceE.parent,targetE.parent),
      getMatchScore(sourceE.parent,targetE),
      getMatchScore(sourceE,targetE.parent)
    )
  // Same level match described in Section 4.5.3.2
  if (isNumeric(sourceE) and isNumeric(targetE))
    if (rangeOverlap(sourceE, targetE))
      sameLevelScore = 1
    else
```

```
    sameLevelScore = 0
else
  sameLevelScore = n-gramSim(sourceE, targetE)
// Child level match described in Section 4.5.3.3
childLevelScore = n-gramSim(sourceE.childString,targetE.childString)
// Linear combination of the 3 factors
Return (α*sameLevelScore + β*childLevelScore + γ*parentLevelScore)
```

4.5.3.1 Parent level match

Parent level match refers to the likely quality or similarity of the source and target elements with respect to their structural context based on the elements above these. Intuitively since our method for determining the match score for an element takes into account what would make any two elements a good match; this function can be used recursively to also determine how well parents of the elements in question match. To account for potential differences in granularity such as the source ontology having an additional level of detail about a concept compared to the target ontology, we also examine the match between the source element's parent and the target element, and vice versa for testing if the target ontology has additional granularity. The maximum score of each of these three potential matches becomes the parent level match score. If the element in question does not have a parent (thus it is the root of its ontology) then that match score is taken to be zero. It should be noted that it might be worth further analysis to see if it might make sense if both elements are the roots of their ontologies perhaps the match

should be assumed to be some value greater than 0, but for the experiments in this section 0 was used exclusively.

4.5.3.2 Same level match

Same level match refers to the similarity of the source and target elements based strictly on the elements themselves without respect for any elements outside of the elements being compared. For this match we use a two different comparisons depending on if the elements are determined to represent a number or range. If both elements are determined to represent a numeric, then range alignment is applied; otherwise textual similarity is used to determine element similarity.

Numeric detection

To detect elements that represent a number or numeric range, we use a set of heuristics to avoid a requirement of any a priori knowledge of the form such an element may take for any particular element or ontology. The first aspect of this heuristic is that if words "one", "two",..., "ten" appear in an element, it is numeric. Second, it is also a numeric if only a number appears as the element text. Finally, if a number appears in the text surrounded by spaces it is likely a number (rather than a coded value). Thus for any element this simple set of heuristics can be applied and it can be determined whether the element should be potentially treated as a number rather than strictly text.

Textual similarity

For textual similarity, n-gram similarity was chosen over word analysis due to the vast majority of the metropolitan travel surveys in the archive mentioned above relying heavily

on encoded values and jargon abbreviations that a word based analysis would likely provide limited benefit without providing an extensive dictionary of abbreviations of codes and jargon (outside the scope of this work). The basic idea behind n-gram similarity is that the similarity of two strings can be determined by examining how close the magnitude of counts of sequences of letters appear within each string. Thus, common sequences of letters can increase the similarity of a match even if the "word" as a whole does not match nor are they even real words such as occurs when abbreviations are concatenated. Thus the intuition behind selecting n-gram similarity was that it would be able to note similar sequences of letters that might otherwise be meaningless through word analysis. This hypothesis seems to be supported by our experimental results in Section 4.5.4.

Range alignment

As described above, if both elements being compared are determined to be numeric, rather than textual similarity, range alignment is applied to determine element fit. Specifically heuristics are used to convert the text of the elements to a range and binary matching is applied. If the range of the source element and the target element overlap at all, this is considered a match; otherwise, the two elements are considered to not match.

Extracting the range like the numeric detection process is based on a number of heuristics. The basic assumption for simplicity is that if an element is determined to be numeric then the element must fit into some range representing that element. In other words, once the element has been classified as numeric there is no check to see how well extracting a range out of the

element works to potentially change back to treating it as text if no direct match is found. The pseudo code for the heuristic is as follows:

```
double[] ExtractRange(String elementString)
  double[] range = [0,0]
  range[0] = range[1] = getFirstNumber(elementString)
  // check for second number as end range
  if (hasSecondNumber(elementString))
    range[1] = getSecondNumber(elementString)
  else
    // if contains key word indicating all values below ...
    if (containsMinKeyWord(elementString))
      range[0] = −∞
      // if doesn't contains key words indicating inclusive range
      if (not inclusiveRange(elementString))
        range[1] = range[1] - MINIMUM_VALUE
    // if contains key word indicating all values above ...
    else if(containsMaxKeyWord(elementString))
      range[1] = ∞
      // if doesn't contains key words indicating inclusive range
      if (not inclusiveRange(elementString))
        range[0] = range[0] + MINIMUM_VALUE
```

```
    else
      // range is a single number
return range
```

From an implementation perspective, the functions for extracting the first number and checking for and extracting the second number can be implemented via regular expressions. The list of minimum words was: "LESS", "FEW", "BELOW", "LOW", "SHORT", "UP TO"; and the list of maximum words was: "MORE","GREAT","HIGH","ABOVE","DOWN TO","TALL". The inclusive range check was simply did the string contain the word "EQUAL" or an equivalent comparative operator (e.g. "<="). Once each element is converted to a range, then if the ranges overlap at all then the same level match score is set to 1 otherwise it is set to 0. It should be noted that while this simplistic binary overlap heuristic is applied, it is easy to envision how this could be extended to further classify the mapping type as equal, subset, superset, etc.

4.5.3.3 Child level match

The final factor, *child level match*, refers to similarity in the elements below the source and target levels. In principle this comparison could be as complicated or as simplistic as desired. For practical reasons (performance and avoiding circular score calculations) the child level score in this section is determined by concerning all elements beneath the source node and target node respectively and comparing the n-gram profile of the concatenated strings.

4.5.4 Experiments

4.5.4.1 Methods

Data

The reference alignments that have been used for past Ontology Alignment Evaluation Initiative (OAEI) competitions to judge ontology alignment methods have focused primarily on determining mapping of ontologies based on being able to extract meaning from words within the ontology elements. As a comparative baseline of general performance, results against the common OAEI data sets are included along with the survey alignment numbers. Specifically we identified approximately sixty surveys recorded in the Metropolitan Travel Survey Archive that all involved encoded elements and ranges of survey ontologies that would need to be aligned before data mining could be performed. From these the 2001 Atlanta survey[1] conducted by the Atlanta Regional Commission and the 2002 Anchorage survey[2] conducted by the Municipality of Anchorage were selected. These two surveys used throughout this work have similarity in survey goals and thus similar data semantically, however the encoding methodologies are significantly different resulting in a challenging problem.

An "expert" reference alignment was created by hand to allow the the quality of any generated alignments to be measured. To minimize bias, the reference alignment that was created was reviewed by an outside person with transportation expertise for accuracy. Below we ex-

---

[1] `http://www.surveyarchive.org/Atlanta/Atlanta2001.xml`

[2] `http://www.surveyarchive.org/Anchorage/Anchorage.xml`

amine how well the mapping algorithms perform on the Atlanta to Anchorage mapping, which we label *travel survey*, to evaluate their performance on encoded and numeric range alignment. While in this section we are primarily interested in encoded and numeric range alignment performance, we also examine alignment performance on other widely used data sets as a measure of traditional alignment performance. Specifically we include results on the commonly used reference alignment datasets of people and pets, Russia, weapons, and networks [1].

4.5.4.2 Experimental evaluation

In the experiments below, the linear weighting of the parent, same-level, and children scores are set to $\alpha = .5$, $\beta = .4$, and $\gamma = .1$. These values were determined through a small set of experiments to determine a good mix of the linear weighting combination. It should be noted that the selection for threshold was selected for each algorithm based on its optimal F-score on the travel survey alignment problem and then this same threshold level used for all other ontologies. Below we refer to the algorithm introduced in this section as the "CWilliams w/Range" algorithm. To allow a fair comparison of what portion of the results can be attributed to the range alignment rather than the choice of n-grams over word analysis, results have also been included with out range alignment ("CWilliams w/o Range"). For comparison we also examine the Descendant's Similarity Inheritance (DSI) algorithm and SSC algorithms as implemented in AgreementMaker as comparative baselines (Sunna and Cruz, 2007). It should be noted that none of the commonly used reference datasets contain numeric values to align

[1] http://oaei.ontologymatching.org/

much less numeric ranges (which was the primary reason the technique introduced here was created). Because none of these data sets contain this feature, the purpose for showing these results is to demonstrate that while the proposed algorithm has been specifically designed for code values and numeric ranges, it still performs relatively well.

As Figure 21 shows although the isNumeric heuristics are fairly simple, adding the numeric heuristic has minimal impact on performance when ontologies do not contain numeric ranges yet do provide significant benefit for the survey ontology that does. It should be noted that in fact for only one data set, Russia, did it degrade the performance at all; precision was 35.6% without range alignment and 35.3% with range alignment, however this difference was not statistically significant. Figure 22 illustrates the recall of each of the algorithms on each of the test data sets. As these results demonstrate, the recall of the proposed algorithm is comparable for all of the data sets and significantly better for the travel survey data set which contains numeric ranges. Finally, Figure 23 shows the F-measure of the algorithms which can be interpreted as the combined performance of each algorithm. As this figure shows the CWilliams algorithms do slightly better on some data sets and poorer on others, but for the problem we are most interested in, travel surveys, it performs significantly better. It is also worth noting that even though it does perform better at 32.7%, across all of the algorithms, this combination of coded values and numeric ranges is by far the most challenging to align of the alignment problems examined here.

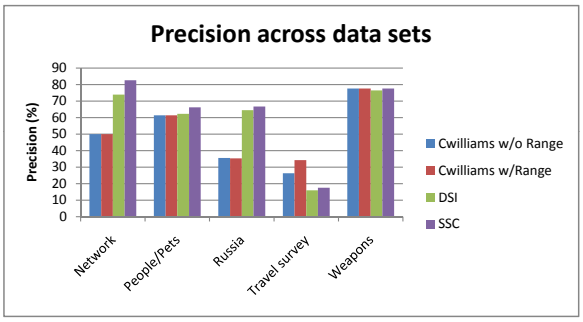

Figure 21. Comparison of precision across various reference data sets.

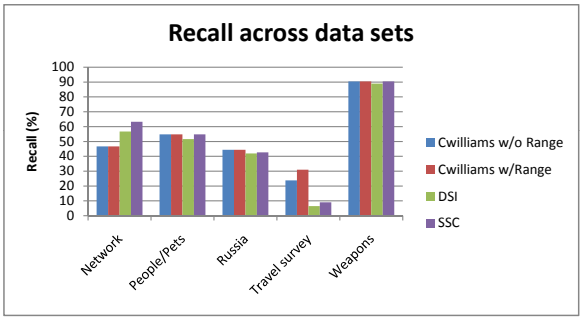

Figure 22. Comparison of recall across various reference data sets.

4.5.5 Discussion

As we have shown in this section, the algorithm proposed here that combines structural alignment ideas with novel techniques for numeric range alignment performs significantly better than previously introduced techniques for survey alignment. However, as the low absolute F-scores reflect, this problem remains a significant challenge for automated alignment. There are a number of refinements to the proposed technique that may offer improvements. For example

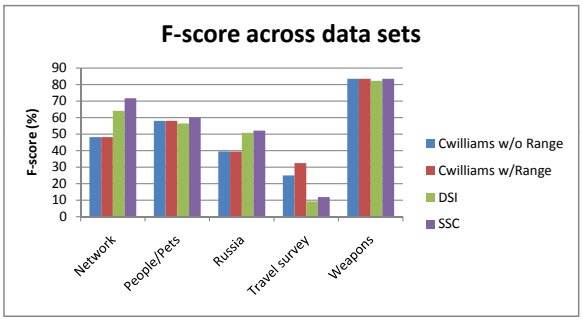

Figure 23. Comparison of F-score across various reference data sets.

in this study no data or hints specific to the travel survey data set were used; however, including such items as a domain specific jargon library or abbreviation database would likely lead to significant improvements. Other items might include more sophisticated numeric checking to avoid misclassifications or a combined n-gram and word based approach to gain the benefit of n-grams for non-words and word-based approaches when possible.

The final area for significant improvement is performance. Currently all source node/target node combinations are tried in agreement maker, a more efficient way to do this for the algorithm listed here would be to use a priority queue and only try matching with children of the top N parent match scores. While limiting the matching to just the top N does potentially have the draw back of missing a better match that has a parent at $N + 1$ or greater distance from the source element's parent. However with a quick test of this using $N = 5$ for the travel surveys showed that the same F-measure could be obtained much faster using this heuristic. While this was a very small test and this point would need to be explored further before any significant conclusions could be drawn, the preliminary results look promising. Also, given that the final

combined score is a linear combination rather than using a top N approach the parent scores of the priority queue could be used to establish a bound where it could be proven algebraically that no parent with a score below a certain point could possible have a score higher than the highest score found so far. Thus the best potential matches could be tried first and the possible matches could be terminated after all potential matches with a better score than the current score had been exhausted.

4.6 Conceptual model

Above the various components that may be used for activity pattern prediction have been introduced and discussed individually. As discussed, each of these components address different aspects of the overall goal of learning the activity patterns of an individual. This section begins by giving an overview of why a conceptual model for individual activity pattern prediction is warranted. Next, a conceptual model is introduced that establishes a general framework of how activity patterns of an individual may be learned. The various components that have been introduced as part of this model and the benefits of addressing each of these areas is discussed as well as how this work fits within this model. This is followed by a discussion of how while this framework encapsulates the various components of this work, the intent of this model is to establish how this problem may be approached in general. As will be discussed, the structure of this model allows improvements to be made to any component of the model without impacting the overall conceptual model. As such, the model introduced in this section can provide a general basis for any future work in this area of study.

4.6.1 Introduction

Throughout this chapter a number of different techniques have been introduced as ways to help either directly with activity prediction or related to ways of either improving or simplifying different aspects of activity pattern prediction. The purpose behind this multi-pronged approach is to acknowledge that there are multiple components that can be used to improve individual activity pattern prediction. Learning the behavior of an individual is very complex when considering the number of spur of the moment decisions and patterns evolving over time whether slowly from changes in habit or abruptly from change in job. As a result, a single learning approach while it may suit one situation well may not be the best approach for other scenarios. Below a conceptual model is introduced that is intended to be flexible enough that multiple approaches to this complex task can explored while still fitting into the proposed framework.

4.6.2 Conceptual model

In developing a conceptual model for the task of individual activity pattern prediction, there were two primary goals. Create a model that is flexible enough to not only capture the techniques introduced in this work, but also fit with approaches explored by others. Second, the model should be simple enough that it is easily understandable and approachable. By addressing these two goals the hope is that the model introduced here can be easily adopted by other researchers going forward.

The proposed conceptual model is shown in Figure 24. The model is made up of three general goals: learning the patterns of the individual, augmenting these patterns from outside

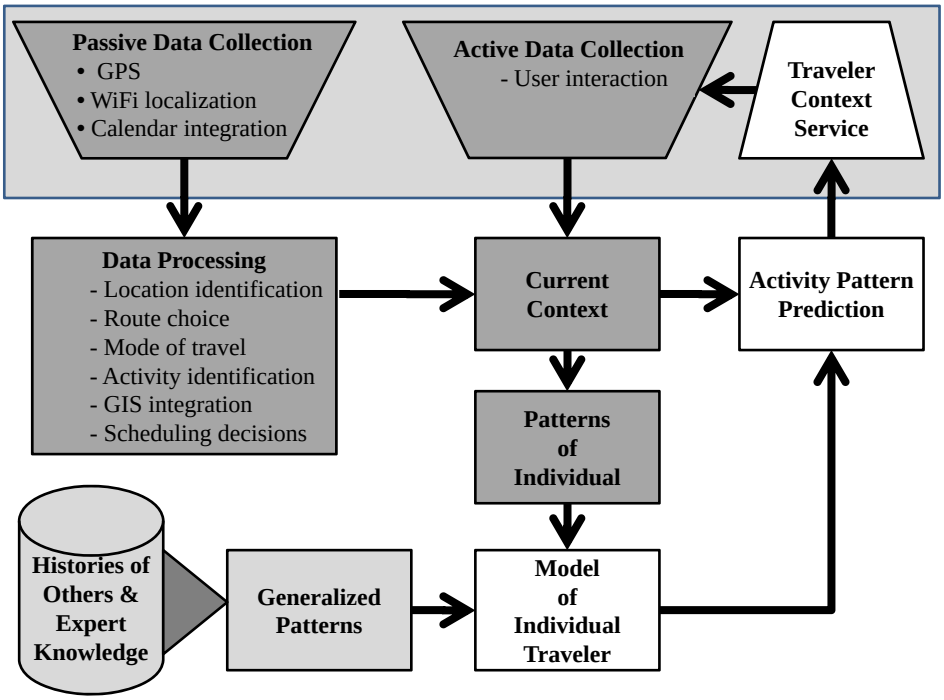

Figure 24. Conceptual model of learning traveler activity patterns.

sources, and developing the individual traveler context from their activity patterns. Each of these goals is addressed through a series of components related to the task discussed further below. The components of the task of learning the patterns of the individual is portrayed in dark gray within the figure and addresses data collection and processing related to learning patterns specific to the individual being observed. The second goal, whose components are displayed in light gray, consists of leveraging outside knowledge or observations and any related transformation necessary to allow this information to augment the data specific to the individual. Finally, the two previous goals are brought together to accomplish the task of predicting

traveler context from the learned activity patterns. The components related to this task, shown in white, include integrating the patterns of the individual and outside knowledge to produce a model of the individual traveler and how that model can be used ultimately to provide and project the traveler context of the individual. A further description of these components is given below.

4.6.3 Learning patterns of the individual

The first of these areas, learning the patterns of the individual, is made up of five components: passive data collection, active data collection, data processing, current context, and patterns of the individual. These components together provide the tools for collecting observations of the traveler and assembling these observations into a model of the individual traveler. The combination of passive data collection plus data processing provide the implicit context of the traveler which can then be augmented or overridden by data that is actively collected from the individual. This combined view of the current context can then be used to help further refine the model of that individual traveler. Section 4.2 touched upon how all of these components can be integrated to form a fluid system for traveler data collection. However, as integration of these sources is not required a more in depth description of the traits of each of these components is given below to provide the flexibility to allow this model to be more generalized.

4.6.3.1 Passive data collection

Passive data collection represents any source of information that can help provide data about the user's current context or future plans that don't require interaction with the user. The gen-

eral requirement of these sources is that the data can be obtained unobtrusively such as carrying a device with them. The benefits of GPS data collection has been a central theme throughout much of the research in this work, but other sources may also provide useful information. One example of this is the use of WiFi networks to either identify the general area of a person or even a good approximation of someone's location indoors based on signal strength (Kawaguchi, 2009). Other studies have shown that cell phone triangulation works quite well at identifying a person's rough location outdoors as well as indoors (Varshavsky et al., 2006).

In addition to senors that observe location, there are a number of other passive sources that might also help establish an individual's context. For example Azizyan and Choudhury have shown that even sensing ambient light and sound can often be suffienct for detecting the type of environment the person (Azizyan and Choudhury, 2009). Besides traditional sensors, there is also an opportunity to passively collect data from sources the person has actively entered data but not specifically for the purpose of the system. An example of this would be a calendar application which the user already keeps track of various events and/or meetings. Integrating of this data with other sources, could allow activities to be tied to calendar events for even more detail without interaction from the user.

4.6.3.2 Active data collection

The second component of data collection is active data collection. The purpose of this component is to collect data that might help establish the context that can not be determined from available passive sources. Like the various types of passive sources, this component is not required, but having some sort of interaction can help capture data that would likely not be

captured by passive sources alone. The main drawback of active collection is it directly impacts the effort required of the individual to get benefit out of the service. Thus the goal would be active collection would take more of an optional role; whereby if the data is provided it helps, but is not required to make the application function. By having the input be optional, the result is a data set that may have missing data from when a user either opts not to answer or was not prompted to reduce their burden. It is in this role that techniques, such as the ACR algorithm introduced in Section 4.3, serve a critical role in minimizing the impact of missing data on the resulting prediction model.

While the user interaction may not require any additional information, more sophisticated techniques discussed in this and other works use knowledge of the current context to determine the active data collection needs. One form of active collection discussed in this work is prompting the individual for planning information or activity information only when the information can not be determined from passive sources alone. In our study collecting additional data was the primary focus of the active user interaction, however other studies have taken different approaches. Liao *et al.*'s application "Opportunity Knocks" observed travel patterns and would prompt user's when they deviated from the known patterns if that was their intention, to help cognitively-impaired people use public transit safely (Liao et al., 2007). As can be seen by applications such as these, while any of these questions could be asked at any point by using the current context, intelligence can be incorporated such that questions only be asked when certain conditions are met as a way of reducing the burden on the individual. Through interac-

tions such as these it is easy to see how occasional interacting with the user might significantly help set the context of the user beyond what can be collected without their interaction.

4.6.3.3 Data processing

The data processing component is responsible for taking passively collected data and synthesizing it into a form that can be used to help provide the individual's context. Within this work we introduced ways GPS data can be processed as it is collected to identify significant locations, returns to locations of interest, and start and stop of travel. These are just a few of the ways passive sources can be processed to provide context. Techniques are being explored for processing GPS logs for mode detection and locations of change of mode as well (Schuessler and Axhausen, 2009). Other work has focused on future end-to-end route prediction based the individual's history and recent GPS trace (Froehlich and Krumm, 2008). Other work in ubiquitous computing has examined activity identification as a function of visiting known locations (Abowd et al., 1997). Similar to their work, one could imagine a similar approach being used matching up an individual's current position with detailed geographic information system (GIS) maps to determine the activity on a larger scale.

In addition to these direct approaches to processing the current location information, other approaches have been used that combine location information with other sources such as calendars. By integrating these different types of passive sources such as the user's calendar and when various events were entered or rescheduled combined with the GPS information can provide additional context of whether the current activity was scheduled or whether the current activity implies rescheduling other activities (Clark and Doherty, 2008). Thus while the data

processing component may take many different forms, the overall goal of these different methods is to enhance the raw passively collected information into a detailed description of the individual's current context.

4.6.3.4 Current context

The component referred to as current context represents the synthesis of all information either through data processing related to passive data sources or from active sources into a single view of what is currently observed about the individual. This snapshot of the context of the individual serves two roles. First the current context can be used to help further build a model of that traveler's individual history or patterns. Second, the current context can be fed into the prediction model to tailor the predictions based on the current knowledge of the state of the system.

4.6.3.5 Patterns of individual

The patterns of the individual represents the history of activity that is collected about that specific individual. This history can take many different forms. Within this work the traveler history or patterns of the individual we have focused on the frequent sequence patterns of traveler context. While this is well suited for the goals of this work namely activity context prediction, there are many other aspects of activity pattern predictions that require a different form of information. For applications that have looked primarily at route prediction, this history is primarily past movement patterns. Scheduling history or how the individual schedules specific tasks and makes adjustments to their schedule as conflicts arise would be another pattern

of interest. The combination of these varying types of history or patterns collected on the individual represent the sum of the knowledge that is specific to that individual.

4.6.4 Outside patterns

One of the challenges in learning the behavior of an individual is that it can take a long time before all intricacies of their behavioral have been observed. If only the individual's behavior is used the result is that it may take a lengthy period of time before the model can be useful. A similar problem occurs frequently when a system first starts to learn the patterns of something with very little history, referred to as the cold start problem. One of the ways this is addressed is by incorporating common patterns outside of the history of the individual to jump start the learning process until more observations are made. Within the conceptual model this process is split into two components: histories of others and expert knowledge; and generalized patterns.

4.6.4.1 Histories of others and expert knowledge

The various sources of outside patterns that might be incorporated into the learning or modeling process of the activity patterns of the individual is represented by the histories of others and expert knowledge component. Traditionally the sources used for helping identify typical patterns are derived from either the history observed in other users or expert knowledge. An important aspect to consider about these sources is while they may be in an identical format as the patterns recorded for the individual being modeled, this does not have to be the case. As shown in this work, similar patterns can be drawn from histories in the same format, but it is also possible for outside sources to be very different in their raw form. This component

represents not only the various sources of the patterns but also any transformation necessary for the sources to be meaningfully incorporated in the learning process.

This work demonstrated that histories could be either used from the same format or outside histories recorded in different could also be used to help aid in rapidly producing a meaningful model of the traveler until a larger base of training examples specific to the user could be collected. As demonstrated in Section 4.5, by mapping the underlying concepts or meta data to the format desired, the source of these histories can vary dramatically in terms of differences in location and transit networks. Despite these differences we have shown that behavioral patterns can still be beneficial in the absence of additional data specific to the individual or even examples from their own city.

While in this work we have focused primarily on showing the transference of activity patterns to aid in learning, this component also may represent other ways histories may be beneficial in learning about the patterns of the individual. Ashbrook and Starner demonstrated that this basic idea could be used within similar groups of people to label significant locations, which would be one way of learning the activity patterns of someone without them specifically entering information about a place (Ashbrook and Starner, 2003). Perhaps the most common forms of this is traffic conditions where the anticipated travel travel for a user does not have to rely on that user's history alone but the travel times of others can help predict the expected travel time of the individual in question.

In addition to using observations of others as a way to quickly form a basis of typical patterns, another source for these patterns can be in the form of expert knowledge. An example

of this would be estimating public transit times. Given the extensive number of public transit routes within a large city, even with a large set of similar histories in the area of the person it would be possible to not have complete coverage of the network. One way of addressing this problem would be to use knowledge of the transit agency to populate the data on the public transit network and schedules. As these various examples demonstrate the use of the histories of others and expert knowledge can serve a critical role in augmenting the history collected specific to the individual.

4.6.4.2 Generalized patterns

Once outside sources have been identified and transformed into a form that is meaningful for expanding on the patterns of the individual, there may need to be some additional form of processing needed. The generalized patterns component represents any processing necessary to allow these various sources to form generalized patterns that can be used in the absence of data specific to the the traveler. As discussed in Section 4.4 there are multiple ways outside sources can be mined to form a generalized set of activity patterns. In this fashion, these patterns can form a collection of typical patterns that can be drawn from to model patterns beyond that observed in a limited history of the traveler.

4.6.5 Developing traveler context

Once the patterns of the individual and any outside patterns have been collected, the task becomes integrating these different views (if both are used) to complete the current traveler context and project future context. The following conceptual components are proposed for

represent this process: model of the individual traveler; activity pattern prediction; and traveler context service. A description of each of these components is given below.

4.6.5.1 Model of individual traveler

The model of the individual component represents how the combination of individual patterns and generic patterns are combined to create a consolidated model of person. Within Section 4.4 we described one way the patterns of the individual could be combined with common patterns extracted from the history of others. As this work demonstrated, by using a combination of individual and general patterns a model can be created that more rapidly produces meaningful prediction than using the patterns of the individual alone. The method introduced in this work is one way that this type of data may be combined, however this model should be general enough that other techniques of combining observations of the individual with other patterns may create a better model than that of either alone.

4.6.5.2 Activity pattern prediction

The activity pattern prediction component is responsible for taking a combined model of the individual traveler and creating predictions of the activity patterns of that person. Specifically this represents how the current observed context is used in conjunction with the model of the traveler to augment the the information observed in the current context. This may take the form of inferring the activity taking place, how flexible the duration of the activity is and whether the current activity was planned or likely a spur of the moment decision. In addition to completing the current context as well as the likely immediately next context, reasonable projections of multiple steps in the future is also possible. While this work has focused primarily on completing

and predicting activity patterns, other forms this can take in enhancing the activity pattern of the user might include identifying their likely route given recent observations. Thus, this component is simple enough to describe the process of predicting elements of current context as well future contexts while still being flexible enough to include behavioral information as well as physical travel details.

4.6.5.3 Traveler context service

The traveler context service is intended to represent the output of the learning and predicting individual activity pattern prediction system. Unlike previous model of individual prediction based primarily on movement patterns, the purpose of the component represented by this service is to provide a combination of not just movement projection but the behavioral activity pattern context. Essentially this component provides this broader context in completing the current context as well as projections of future contexts. Within this work we have primarily addressed adding the new element of activity and behavioral information to the traveler context, this concept of context could easily be expanded to include preferred routes, household constraints or even the affect of pricing on behavior. By addressing this larger picture of the likely future behavior and associated constraints this service could provide a wider range of information that could be useful for mobile applications.

4.6.6 Discussion

As this section has shown this conceptual model represents a consolidated way addressing the various components of individual activity patterns and the associated traveler context. Although this model has been examined with respect to the parts of the research in this work,

the conceptual integration of these components is flexible enough that the hope is it will provide a standard framework for other studies in this area. In describing this model the focus has been on explaining how each of these different components can fit together to address the larger goal of activity pattern projection, however it is important to note that solutions could just address portions of this overall model without impacting the representative value of the model. For example, models discussed earlier that use individual history alone would just not include the components related to incorporating the histories of others or expert patterns while the remainder of the model would still describe the work. Likewise, models of generalized behavior to simulate the activity patterns of individuals would also fit this overall model just without the incorporation of individual patterns. As a result, we propose that the model introduced here could for a basis for describing and evaluating individual activity pattern prediction going forward.

CHAPTER 5

SUMMARY AND DISCUSSION

Prior to this work, considerable research had been made in trying to predict where a specific individual was likely to go in the future, however the reason for the travel and any associated constraints were largely overlooked. Since the majority of these approaches relied on the location traces of the individual, an additional weaknesses of these approaches was a lengthy history of observations was necessary and predictions could only be made for previously visited locations. Complementary work had taken a different approach examining the behavior of travelers at an aggregate level to better understand and simulate travel behavior in general but the models were not suited to modeling the behavior of an individual. The purpose of this work was to address this gap and develop a better model of individual behavior.

5.1 Statement of problem

The problem this work addressed was developing a method for rapidly learning traveler activity patterns for mobile applications. Accomplishing this meant addressing several goals. First, developing a way to predict the behavioral reasons behind travel specific to an individual was needed. Furthermore, the method should flexible enough not to be geographic specific both in terms of being city specific and being able to infer this behavioral information even if the destination had not been previously visited. Finally, to address the needs of mobile applications two aspects had to be addressed. First, a reasonable model of the individual's behavior needed

to be developed quickly without requiring a huge data entry burden on the traveler for an extended period of time. Second, the influences on the behavior of the individual should be understandable such that mobile applications could take action to influence behavior.

5.2 Review of methodology

Using a quantitative research design, we have demonstrated the findings of this work in an empirical fashion. This design was chosen as it best suited the task of data mining and prediction of known variables. In addition to the quantitative measures used in evaluating predictive performance, qualitative comparisons were also used for comparing the descriptive power of the models generated and their ability to be leveraged by other applications.

To address the task of learning the activity patterns of individuals, a subset of attributes typically used in travel behavior activity surveys was used. The attributes that were selected capture activity, location, mode of transportation, arrival time, departure time, duration, sequential order, and planning flexibility information of the activity and travel of the person. These attributes were selected as they have been shown to effectively predict the travel and activity patterns by transportation planners at an aggregate level while they capture the information of interest in modeling an individual traveler.

These traits were discretized and the series of activities and associated travel of the participant became a series of sets representing the events or traveler context throughout their schedule. To demonstrate the findings of this study three transportation surveys collected outside of this study in addition to a collection effort conducted as part of this work were used. Due to differences between these studies, the data captured in the studies had to be aligned by

hand such that the patterns identified in one survey could be meaningfully examined in relation to the other studies. To fit this same model the GPS trace that was collected as part of this work was processed to automatically identify discretized versions of the attributes matching the study's goals.

Using this approach, the prediction problem could be described as given a series of traveler contexts gathered through active and passive means, predict the next set or sets of contexts or missing attributes of interest in the current or previous contexts. Information retrieval measures were then used to quantitatively analyze the value of the predictions of the model created. To demonstrate the soundness of this work; the primary goal was broken down into a number of sub-problems. First, the internal validity of being able to predict the activity patterns of a traveler and generation of a traveler schedule with "perfect" information was shown. Next, techniques are introduced and analyzed for reducing the burden on the traveler for collecting data. This is followed by presenting algorithms and techniques that can further reduce the data collection requirements and thus user burden with limited impact on the quality of predictions. Afterward, a demonstration of how activity patterns might be transferred from other locations to improve predictions is shown. Finally, a conceptual model of how all of these concepts integrate together for rapidly learning the activity patterns of an individual is presented.

5.3 Summary of results

As this work has shown, through a data mining approach combined with an activity modeling approach it is possible to rapidly learn the activity patterns of individuals in a way that can be leveraged by mobile applications. Prior to this work techniques for individual travel

prediction had focused almost solely on observations of an individual traveler and location and route prediction. Identifying the reasons for the travel and the flexibility of plans were largely overlooked. As this work has demonstrated, by applying an approach inspired by transportation planning referred to as activity-based modeling, a richer model of the activity patterns of an individual could be developed. Furthermore, by using a model that is independent from physical location many of the limitations of previous approaches related to only being able to work effectively with previously visited locations were able to be avoided, resulting in a more robust model that could provide traveler context even when a location had not been previously visited. Furthermore, through leveraging passive data sources, such as GPS, combined with limited user interactions a model suitable for mobile applications was developed that limited user burden and provided an understandable model that could potentially be used to influence the traveler's behavior.

To solve this problem a multi-faceted approach was introduced that demonstrated several ways of improving activity-based predictions. First, this work demonstrated the predictive benefits of using associative sequence rules to predicting activity sequences over techniques such as Markov and Bayesian models. This work demonstrated activity-based approaches that had previously been only validated against aggregate activity distributions could also be tailored to predict activity patterns of specific individual. This was demonstrated both in determining the next step in a travel sequence in addition to making reasonable predictions for multiple steps into the future as shown in Section 4.1.

Since traveler burden is a major concern for practicality for mobile applications, the techniques introduced in this work for reducing participant burden and improving predictions in light of missing data are one of the most significant contributions of this work. Addressing this took several techniques. First, new methods were introduced for processing GPS streams to identify and match significant locations were introduced. The accuracy of this method and the activity prediction methods was demonstrated with a large survey that was conducted as a part of this research. Since it was impractical to ask extensive questions from users on an ongoing basis, techniques that could better address the problem of missing data in sequences were needed. A primary contribution of this work was a technique that greatly reduced the negative impact of missing data on sequential associative prediction. As the results demonstrate, this technique is far more robust to even large amounts of missing values with limited impact to predictive quality.

One of the weaknesses of prior approaches has been a lengthy history of the individual's travel was required before predictions could be made. To address this challenge so that meaningful predictions could be made far more rapidly, a technique was introduced for allowing the travel patterns to be transferred across physical locations so that the histories of others could be used to jump start the predictions of the individual until a more extended history of the individual could be collected. This study both introduced how the meta data of activity patterns of one city could be used to predict the activity patterns of another city. In addition to this technique, a algorithm was introduced that could reduce the time required to create

mappings of activity data by automatically identifying the mappings of many concepts across activity-based surveys recorded in multiple formats.

Finally, a comprehensive conceptual model was introduced for the general task of individual activity pattern projection. The introduced framework identifies components in detail that are both understandable and flexible. This model is intended to be flexible enough to apply to multiple different approaches, thus providing a model for other studies in this area going forward.

5.4 Limitations and boundaries of study

While this work represents significant progress toward learning the travel and activity patterns of individuals, it is only a first step in what may be possible and it does have its limitations. For one, this study has focused only on urban areas within North America. An interesting area for future research would be to explore at what population size does the transferability of urban patterns to more rural areas breakdown assuming that this does occur. Another area worth further study would be studying how this work could be used internationally. While Timmermans et al. have demonstrated while some activity patterns can be transferred across different cultures at an aggregate level, it would be interesting if this holds true at an individual level (Timmermans et al., 2003). Also, although this study has demonstrated learning patterns of individuals in North America; there is nothing within the learning itself that is specific to patterns observed in that area. Therefore the techniques introduced in this work would likely also be successful in modeling the patterns of individuals in other countries as well although this would need to be demonstrated.

Another current limitation of this research is the dependency on needing active user input compared to the completely passive model that have been used previously in ubiquitous computing. The amount of effort a mobile user is willing to spend on active entry to gain the benefits of a more customized model remains to be seen. A key factor that could encourage users into being willing to provide more information would be if applications are able to provide useful personalization based on the amount of information known about the user. Another aspect that could help reduce the need for user input is more extensive labeling of locations. As more and more people use GPS enabled devices, by sharing their labeling of locations it would be possible to extract some level of detail about what the user might be doing based on the activities of others at that location.

Finally, due to the associative mining being resource intensive both computationally as well as the memory requirements, it probably would not be practical to perform the data mining on current mobile devices. As a result, one potential way of handling this would be wireless connectivity with a server for processing and having just the resulting model sent back to the mobile device. This would allow the model to be mined on a more suitable platform, while still being able to make timely predictions from a local model.

5.5 Discussion of results

As this work has shown, there is a considerable amount of information that can be learned about a traveler's activities and their plans beyond just where they are likely going. With the projected increase in market penetration of GPS-enabled personal data assistants (PDA) and cell phones (Research, 2009; iSuppli, 2007), the market is ripe for the emerging fields of location-

aware and activity-aware applications. Potential applications in this space could enable users to get recommendations that are personalized to match the user's observed preferences. Through being able to identify a user's activity patterns and the factors influencing these behaviors, a new generation of intelligent traveler's assistant applications can be possible as well as other applications. With insights such as where a person is likely to be eating in the near future, it is easy to envision how an application might use this information to display new restaurants in the area or promote a restaurant in the area that has not been tried before. Other uses might include using their mode preference information to either display traffic conditions or pull public transit information to alert the traveler of when the next bus is going to arrive going to their next destination. In fact, Gartner projects that money spent on mobile ads will surpass 13 billion by 2013, up 248% from 2008 (Gartner, 2009).

One of the open questions in this area is how accurate do predictions need to be for various applications. For mobile applications there are likely some predictions that a low level of precision would be required to still provide value such as when to retrieve nearby public transit information as there is little penalty if the prediction is wrong. However, if the application is constantly suggesting different locations for activities that the user is not interested in this would provide little value and might even be annoying turning user's away from the application. Possible ways of handling or learning what type of information the user is interested in and not interested in might be to use current techniques based on positive and unlabeled learning where interaction or following suggestions would be positive examples and non-response would be unlabeled examples (Li and Liu, 2005). For other uses such as imputing missing values for travel

surveys or trying to model micro-simulations more accurately increasing the precision at the cost of recall would likely be preferred as data quality would be a much higher concern. While ways to balance the precision and recall to a specific application has been shown, determining what is the best mix will likely be specific to the targeted use. In addition, as this is a first step in this direction, it may turn out that a higher degree of precision than this initial model delivers may in fact be necessary for some applications. Determining the quality of prediction for each of these different applications should be a focus of future research in this area.

Another aspect of this work that deserves further exploration is determining the best combination of active and passive learning. While additional active information would almost always help, as it would provide continual supervised learning essentially, the cost in user time would more than likely not be worth the additional benefit. While the exact mix of the amount of actively collected data needed to ensure a desired level of precision would vary depending on the target uses, discovering the relationship between the predictive benefit versus the timesavings would be of great interest. Between transportation activity survey design and mobile applications, the desired level of precision would likely vary, but determining the time cost versus payoff curve would yield valuable information for future work in this area. Also along these lines this trade off would also be interesting as additional passive data sources are introduced and how they affect the cost benefit relationship.

Personalization has proven quite valuable on many internet sites, but this type of capability to personalize users' experiences based on their travel patterns remains an open problem. Two key problems must be solved before these types of applications can become a reality. The first

challenge is how to best use the patterns of one person to benefit another. The second is how to ensure that a user's movements and behaviors are protected in order to prevent direct effects on their privacy and physical security. As the world wide web has shown, collaborative filtering has many benefits of improving user's experience through leveraging the histories of others. While much of this type of personalization on the web has limited impact on compromising user privacy, when dealing with a user's travel history and frequented locations this represents a significant threat to a person's privacy or even security. As this work has shown a meaningful model can be built of the user while still protecting their privacy, but there will also likely be many benefits to be gained by also including the experiences of others. Developing privacy preserving collaborative filtering offers the potential for huge benefits.

Specifically there is likely a great amount of benefit to be gained by further reducing the information required from the user by using information collected from other users. While traveler personalization may seem straightforward, there are some significant differences between a recommender for a traveler compared to a recommender for an internet user. Ideas that remain the same include shared preferences and common likes/dislikes; however, with travelers there are additional considerations as well. In particular, a key challenge for traveler recommendations is being able to account for spatial and time constraints. For example, while two different people may visit the same restaurant, one person may reside 5 minutes from the restaurant, while the other may reside an hour away. As a result, the frequency of recommendations in the area of that restaurant are likely very different for each user. A similar problem occurs if users' time constraints vary. For example, imagine two users have similar transit preferences, for one

user the train is the preferred way for traveling between points A and B, but for the other user while the train might be the preferred option, tighter time constraints may prevent taking the train from being a viable alternative.

The goal of such work would be collaborative recommendation systems for travelers such as those currently available for web applications, which preserve the privacy of the user. This type of application could have both commercial and personal benefits (Macker and Corson, 1998). For example, a restaurant located in an area where a set of users are likely to be during lunch time could push coupons or ads to this target set of users so that it would likely have a much higher conversion rate than when traditional ads are used. For users, possible applications could include additional locations for a desired activity, best matching their location and mode of transit preferences. Other applications might include being able to query how many people frequent each restaurant in a given area providing users with a different way of deciding where to go rather than relying on reviews alone. From these applications, it also becomes apparent that malicious manipulation of these recommendations could potentially be very profitable, while simultaneously reducing users' trust in the system. To summarize, efforts to create recommender systems for travelers involve two main sub-problems. First, in order to create a meaningful collaborative traveler recommender system, the best way to make use of the histories of multiple users' histories, each of whom may have different motives for going to a particular place or event, would need to be determined. Second, a method for sharing users travel histories for these sophisticated models, that does not infringe on people's privacy must be developed that also protects the system from malicious manipulation.

In addition, one of the key contributions of this work is techniques to effectively learn despite missing data. While this work has primarily focused on learning activity patterns given a data source with many different missing elements, the algorithm introduced in this work is not specific to this task. One area of study would be to explore what other applications may benefit from this technique of learning. Potential applications may include sensor networks where loss of connectivity or faulty sensors may result in periods of missing data. In this situation an algorithm such as ACR may be of great use in inferring the missing elements if needed.

Finally, while much of this work has focused on the applications of this research to mobile applications, this research may benefit transportation planning as well. As this work has demonstrated with some care, activity patterns can be transferred across cities; one area of future research would be to determine how best to take advantage of this to improve activity projects when limited data is available or improve the survey process itself. As this work has shown activity patterns of an individual can be learned quickly when aided by a base of common activity patterns. This combined with being able to greatly reduce the participant burden while still being able to learn patterns may make more extended activity surveys more palatable. Another area that is worth further investigation is how the demonstrated transferability of activity patterns might be leveraged to potentially reduce the size of future activity surveys knowing that the collected data may be augmented by outside data. Finally, as this work has shown, it is possible to build a more accurate model of individual behavior than existing techniques. Utilizing the techniques introduced in this work may also help improve the quality of micro simulations at being able to project behavior that is more realistic.

APPENDIX

DATA SET INFORMATION

This chapter contains the details of the data sets used in this work.

A.1 Chase data set details

A.1.1 Categories of activity types

- Individual mandatory activities (i.e. not conducted with other household members, but may be with other non-household members)
- Joint maintenance activities (i.e. conducted with other household members)
- Joint discretionary activities (i.e. conducted with other household members)
- Allocated maintenance activities (i.e. not conducted with other household members, but may be with other non-household members)
- Individual discretionary activities (i.e. not conducted with other household members, but may be with other non-household members)

A.1.2 High level activity categories

- Night sleep, other needs
- Meals
- Work/School
- Household Obligations
- Drop-off/Pick-up
- Shopping
- Services

APPENDIX (Continued)

- Active recreation
- Entertainment
- Social
- Other

APPENDIX (Continued)

A.1.3 Detailed activity categories

Night sleep	Other household obligations	Medical/professional
Wash/dress/pack/snacks	Attending to pets	Barber/salon/beauty
In-home meal	People	Banking
Bagged lunch	Meal	Religious
Restaurants	Snacks/drinks	Gas
Coffee/snack shop	Video rental	Other service
Other basic needs	Dry cleaning	Hobbies
At work	Mail	Exercise or active sports
Telework	Other items (Dry cleaning, Mail, etc.)	Spectator Events/Theatre
Volunteer work	Convenience store	Playing/parks
At School	Minor groceries (¡10 items)	Regular TV programs
Schoolwork	Major groceries (10+ items)	Unspecific TV
Training/special classes	Housewares	Watching video
Other work/school	Clothing/personal items	Relaxing/ napping/ reading
Cleaning/Maintenance	Drug store	Email/internet
Meal preparation	Internet shopping	Other recreation/entertainment
Attending to children	Other shopping	Hosting visitors
Visiting	Telephone ¿10 minutes	With babysitter
Religious/cultural	Other social	Other
Planned social events	Tag along with parent	Tag along travel
Cultural/recreational/special clubs	Playing, socializing	Filling out this survey
Helping others	Homework	Pleasure driving
Travel coded as an activity		

APPENDIX (Continued)

A.2 Atlanta data set details

The data selected for use in the experiments from the 2001-02 Atlanta Household Travel Survey was all events from the survey that didn't contain any missing values for the attributes: activity, mode of trasportation, arrival time, departure time, duration, and age. This filtering was done to allow the experiments with the amount of missing data conducted in this work to be performed in a more controlled manner. These attributes were broken down into the following discrete values:

- **Activity** - 29 values: 'Other at home activities', 'Working', 'Medical or dental', 'Getting ready', 'Major shopping', 'Personal', 'Watching children', 'Pick up something or drop off something', 'Worship or religious meeting', 'Visit friends or relatives', 'Household work or outdoors work', 'Entertainment', 'Outdoor recreation', 'Fitness or exercising', 'Rest or relax', 'Waiting for transportation', 'No other activities', 'Personal business', 'Other', 'Eating', 'Volunteer work', 'Work related from home', 'Community meetings', 'ATM, banking, post office, bill payment', 'Sleep', 'Work related business', 'School', 'Drop off or pickup someone', 'Incidental shopping'
- **Mode of transportation** - 12 values: 'Walk', 'Auto or van or truck - passenger', 'Other', 'Airplane', 'Intercity train (Amtrak)', 'Transit - MARTA bus', 'Dial a ride or paratransit', 'Intercity bus (greyhound, trailways)', 'School bus', 'Taxi, shuttle bus or limousine', 'Auto or van or truck - driver', 'Motorcyle or moped', 'Bicycle', 'Transit - CCT bus', 'Heavy rail - MARTA'

APPENDIX (Continued)

- **Arrival time** - 8 values: '3am-8am', '8am-10am', '10am-12pm', '12pm-2pm', '2pm-4pm', '4pm-6pm', '6pm-8pm', '8pm-3am'
- **Departure time** - 8 values: '3am-8am', '8am-10am', '10am-12pm', '12pm-2pm', '2pm-4pm', '4pm-6pm', '6pm-8pm', '8pm-3am'
- **Duration** - 7 values: '10 minutes or less', '10-30 minutes', '30-60 minutes', '1-2 hours', '2-4 hours', '4-8 hours', 'Greater than 8 hours'
- **Age** - 9 values: '10 years old or less', '10-20 years old', '20-30 years old', '30-40 years old', '40-50 years old' '50-60 years old', '60-70 years old', '70-80 years old', 'greater than 80 years old'

CITED LITERATURE

Abowd, G., Atkeson, C., Hong, J., Long, S., Kooper, R., and Pinkerton, M.: Cyberguide: A mobile context-aware tour guide. Wireless Networks, 3(5):421–433, 1997.

Agrawal, R., Imielinski, T., and Swami, A.: Mining association rules between sets of items in large databases. SIGMOD Rec., 22(2):207–216, 1993.

Agrawal, R. and Srikant, R.: Mining sequential patterns. In Eleventh International Conference on Data Engineering, eds. P. S. Yu and A. S. P. Chen, pages 3–14, Taipei, Taiwan, 1995. IEEE Computer Society Press.

Algers, S., Eliasson, J., and Mattsson, L.-G.: Is it time to use activity-based urban transport models? A discussion of planning needs and modelling possibilities. The Annals of Regional Science, 39(4):767–789, December 2005.

Arentze, T., Hofman, F., van Mourik, H., and Timmermans, H.: Spatial Transferability of the Albatross Model System: Empirical Evidence from Two Case Studies. Transportation Research Record: Journal of the Transportation Research Board, 1805:1–7, 2002.

Arentze, T. A. and Timmermans, H. J. P.: Representing mental maps and cognitive learning in micro-simulation models of activity-travel choice dynamics. Transportation, 32(4):321–340, July 2005.

Arentze, T., Borgers, A., Hofman, F., Fujii, S., Joh, C., Kikuchi, A., Kitamura, R., Timmermans, H., and der Waerden, P. V.: Rule-based versus utility-maximizing models of activity-travel patterns: A comparison of empirical performance. In Travel Behavior Research: The Leading Edge, ed. D. Hensher, pages 569–584. Pergamon, Amsterdam, Pergamon, 2001.

Arentze, T. and Timmermans, H.: Albatross: A Learning-Based Transportation Oriented Simulation System. In European Institute of Retailing and Services Studies. Eindhoven, 2000.

Arentze, T. and Timmermans, H.: Dynamic Model for Generating Multiday, Multiperson Activity Agendas: Approach and Illustration. In 86th Annual Meeting of the Transportation Research Board, volume 86, 2007.

Arentze, T. A. and Timmermans, H. J.: A need-based model of multi-day, multi-person activity generation. Transportation Research Part B: Methodological, 43(2):251 – 265, 2009. Modeling Household Activity Travel Behavior.

Ashbrook, D. and Starner, T.: Learning significant locations and predicting user movement with GPS. In Wearable Computers, 2002. (ISWC 2002). Proceedings. Sixth International Symposium on, pages 101–108, 2002.

Ashbrook, D. and Starner, T.: Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users. Personal Ubiquitous Computing, 7(5):275–286, 2003.

Ashiru, O., Polak, J. W., and Noland, R. B.: Development and Application of an Activity Based Space-Time Accessibility Measure for Individual Activity Schedules. In Presented at the Annual Meeting of the European Regional Science Association, Jyvaskyla, Finland, 2003.

Ashiru, O., Polak, J. W., and Noland, R. B.: Space-Time User Benefit and Utility Accessibility Measures for Individual Activity Schedules. Transportation Research Record: Journal of the Transportation Research Board, 2003.

Ashiru, O., Polak, J. W., and Noland, R. B.: The Utility of Schedules: A Model of Departure Time and Activity Time Allocation with Application to Individual Activity Scheduling. In 10th International Conference on Travel Behaviour Research, 2003.

Ashiru, O., Polak, J. W., and Noland, R. B.: The Utility of Schedules: Theoretical Model of Departure-Time Choice and Activity-Time Allocation with Application to Individual Activity Schedules. Transportation Research Record: Journal of the Transportation Research Board, 1894:84–98, 2004.

Auld, J., Mohammadian, A., and Doherty, S. T.: Analysis of Activity Conflict Resolution Strategies. In 87th Annual Meeting of the Transportation Research Board, 2008.

Auld, J., Williams, C. A., Mohammadian, A., and Nelson, P. C.: An Automated GPS-Based Prompted Recall Survey With Learning Algorithms. Transportation Letters: The International Journal of Transportation Research, 1(1):59–79, January 2009.

Axhausen, K. and Grling, T.: Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. Transport Reviews, 12(4):323–341, 1992.

Azizyan, M. and Choudhury, R. R.: SurroundSense: mobile phone localization using ambient sound and light. SIGMOBILE Mob. Comput. Commun. Rev., 13(1):69–72, 2009.

Bashir, S., Razzaq, S., Maqbool, U., Tahir, S., and Baig, A. R.: Using Association Rules for Better Treatment of Missing Values. ArXiv e-prints, April 2009.

Bates, J., Polak, J., Jones, P., and Cook, A.: The valuation of reliability for personal travel. Transportation Research Part E, 37(2-3):191–229, 2001.

Berendt, B., Hotho, A., and Stumme, G.: Towards Semantic Web Mining. In Lecture Notes in Computer Science: The Semantic Web - ISWC 2002, volume 2342, pages 264–278. Springer, 2002.

Bhattacharya, A. and Das, S. K.: LeZi-update: an information-theoretic approach to track mobile users in PCS networks. In MobiCom '99: Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking, pages 1–12, New York, NY, USA, 1999. ACM Press.

Blewitt, G. and Taylor, G.: Mapping Dilution of Precision (MDOP) and map-matched GPS. International Journal of Geographical Information Science, 16(1):55 – 67, 2002.

Bowman, J. L. and Ben-Akiva, M. E.: Activity-based disaggregate travel demand model system with activity schedules. Transportation Research Part A: Policy and Practice, 35(1):1–28, January 2001.

Bricka, S., Zmud, J., Wolf, J., and Freedman, J.: Household Travel Surveys with GPS. Transportation Research Record: Journal of the Transportation Research Board, 2105(1):51–56, December 2009.

Bureau, U. S. C.: Census 2000 - Profiles of General Demographic Characteristics: Illinois. Technical Report D1-D00-DDPR-00-IL1, U. S. Census Bureau, May 2001.

Bureau, U. C.: Annual Estimates of the Population of Metropolitan and Micropolitan Statistical Areas: April 1, 2000 to July 1, 2007. Technical Report CBSA-EST2007-01, U.S. Census Bureau, March 2008.

Canada, S.: Population and dwelling counts, for census metropolitan areas, 2006 and 2001 censuses. Technical report, Statistics Canada, December 2008.

Charypar, D. and Nagel, K.: Generating complete all-day activity plans with genetic algorithms. Transportation, 32(4):369–397, 2005.

Chen, B., Tan, H., and Lambrix, P.: Structure-Based Filtering for Ontology Alignment. In Proceedings of the 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE'06), pages 364–369, Los Alamitos, CA, USA, 2006. IEEE Computer Society.

Clark, A. and Doherty, S.: Use of GPS to automatically track activity rescheduling decisions. In Proceedings of the 8th International Conference on Survey Methods in Transport, 2008.

Cleverdon, C.: Evaluation of Tests of Information Retrieval Systems. Journal of Documentation, 26:55–67, 1970.

Damm, D. and Lerman, S. R.: A theory of activity scheduling behavior. Environment and Planning A, 13(6):703–718, 1981.

Dillenburg, J., Wolfson, O., and Nelson, P.: The Intelligent Travel Assistant. In Proceedings of the IEEE 5th International Conference on Intelligent Transportation Systems, pages 691–696, 2002.

Dillenburg, J. F., Nelson, P. C., Wolfson, O., Yu, O., Sistla, A. P., McNeil, S., Ouksel, A. M., Xu, B., and Ben-Arie, J.: Applications Of A Transportation Information Architecture. 2004 IEEE International Conference on Networking, Sensing and Control, 1:480–485, March 2004.

Doherty, S. and Mohammadian, A.: Application Of Artificial Neural Network Models To Activity Scheduling Time Horizon. Transportation Research Record: Journal of the Transportation Research Board, 1854:43–49, 2003.

Doherty, S. T. and Axhausen, K. W.: The development of a unified modelling framework for the household activity-travel scheduling process. In Traffic and Mobility: Simulation-Economics-Environment, eds. W. Brilon, F. Huber, M. Schreckengerg, and H. Wallentowitz, pages 35–56. Springer, 1999.

Doherty, S. T.: The household activity-travel scheduling process, computerized survey data collection and the development of a unified modelling framework. Doctoral dissertation, University of Toronto, 1998.

Doherty, S. T. and Miller, E. J.: A computerized household activity scheduling survey. Transportation, 27(1):75–97, February 2000.

Doherty, S. T., Miller, E. J., Axhausen, K. W., and Grling, T.: A Conceptual Model of the Weekly Household Activity-Travel Scheduling Process. In Travel Behaviour: Spatial Patterns, Congestion and Modelling, eds. E. Stern, I. Salomon, and P. Bovy, pages 233–264. Edward Elgar Publishing Limited, 2002.

Doherty, S. T. and Mohammadian, A.: The Validity of Using Activity Type to Structure Tour-based Scheduling Models. In Proc. of 86th Annual Meeting of the Transportation Research Board, Washington D.C., January 2007.

Doherty, S. T., Nol, N., Gosselin, M.-L., Sirois, C., Ueno, M., and Theberge, F.: Moving Beyond Observed Outcomes: Integrating Global Positioning Systems and Interactive Computer-Based Travel Behaviour Surveys. In Transportation Research Circular: Personal Travel: The Long and Short of It, pages 449–466, 2001.

Doherty, S. and Miller, E.: Tracing the Household Activity Scheduling Process Using One-Week Computer-Based Survey. Transportation, 27:75–97, 2000.

Doherty, S., Nemeth, E., Roorda, M., and Miller, E.: Design and Assessment of the Toronto Area Computerized Household Activity Scheduling Survey. Journal of the Transportation Research Board, 1894:140–149, 2004.

Eagle, N. and Pentland, A.: Reality mining: sensing complex social systems. Personal and Ubiquitous Computing, 10(4):255–268, May 2006.

Eagle, N. and Pentland, A. S.: Eigenbehaviors: identifying structure in routine. Behavioral Ecology and Sociobiology, 63(7):1057–1066, May 2009.

Edwards, P., Grimnes, G., and Preece, A.: An empirical investigation of learning from the semantic web. In ECML/PKDD, Semantic Web Mining Workshop, volume 14, 2002.

Ettema, D., Bastin, F., Polak, J., and Ashiru, O.: Modelling the joint choice of activity timing and duration. Transportation Research Part A: Policy and Practice, 41(9), 2007.

Ettema, D., Borgers, A., and Timmermans, H.: Simulation model of activity scheduling behavior. Transportation Research Record, pages 1–11, 1993.

Ettema, D., Schwanen, T., and Timmermans, H.: The effect of Location, Mobility and Socio-Demographic Factors on Task and Time Allocation of Households. Transportation: Planning, Policy, Research, Practice, 34(1), 2007.

Euzenat, J. and Valtchev, P.: Similarity-Based Ontology Alignment in OWL-Lite. In Proceedings of the 16th European Conference on Artificial Intelligence (ECAI-04), pages 333–337, 2004.

Facca, F. and Lanzi, P.: Mining interesting knowledge from weblogs: a survey. Data and Knowledge Engineering, 53(3):225–241, 2005.

Flamm, M., Jemelin, C., and Kaufmann, V.: Combining person based GPS tracking and prompted recall interviews for a comprehensive investigation of travel behaviour adaptation processes during life course transitions. In Proceedings of the 7th Swiss Transport Research Conference, Monte Verita, Switzerland, September 2007.

Flotterod, G. and Nagel, K.: Modeling and estimation of combined route and activity location choice. In Proceedings of ITSC 2006, 2006.

Fossati, D., Ghidoni, G., Eugenio, B. D., Cruz, I. F., Xiao, H., and Subba, R.: The Problem of Ontology Alignment on the Web: a First Report. In Proceedings of the 2nd Web as Corpus Workshop of the 11th Conference of the European Chapter of the ACL, April 2006.

Frank, L., Bradley, M., Kavage, S., Chapman, J., and Lawton, T.: Urban form, travel time, and cost relationships with tour complexity and mode choice. Transportation, 35(1):37–54, January 2008.

Frignani, M. Z., Auld, J., Mohammadian, A., Williams, C., and Nelson, P.: Urban Travel Route and Activity Choice Survey (UTRACS): An Internet-Based Prompted Recall Activity Travel Survey using GPS Data. In Proceedings of 89th Annual Meeting of the Transportation Research Board, Washington D.C., January 2010.

Frignani, M. Z., Auld, J., Mohammadian, A., Williams, C., and Nelson, P.: Urban Travel Route and Activity Choice Survey (UTRACS): An Internet-Based Prompted Recall Activity Travel Survey using GPS Data. tentatively accepted for publication in Transportation Research Record, January 2010.

Froehlich, J. and Krumm, J.: Route Prediction from Trip Observations. Technical Report 2008-01-0201, SAE International, Detroit, MI, April 2008.

Frusti, T., Bhat, C., and Axhausen, K.: An exploratory analysis of fixed commitments in individual activity-travel patterns. Transportation Research Record, 1807:101–108, 2003.

Garofalakis, M., Rastogi, R., and Shim, K.: Mining sequential patterns with regular expression constraints. IEEE Transactions on Knowledge and Data Engineering, 14(3):530–552, May/June 2002.

Garofalakis, M. N., Rastogi, R., and Shim, K.: SPIRIT: Sequential Pattern Mining with Regular Expression Constraints. In VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases, pages 223–234, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

Gartner: Mobile Advertising Quietly Grows. http://www.gartner.com/DisplayDocument?id=1147013, August 2009.

Giuliano, G. and Narayan, D.: Another Look at Travel Patterns and Urban Form: The US and Great Britain. Urban Stud, 40(11):2295–2312, October 2003.

Gogate, V., Dechter, R., Bidyuk, B., Marca, J., and Rindt, C.: Modeling Transportation Routines using Hybrid Dynamic Mixed Networks. In Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI), Edinburgh, Scotland, July 2005. AUAI Press.

Grling, T., Kaln, T., Romanus, J., and Selart, M.: Computer Simulation of Household Activity Scheduling. Environment and Planning A, 30:665–679, 1998.

Grling, T., Kwan, M. P., and Golledge, R. G.: Computational-Process Modelling of Household Activity Scheduling. Transportation Research, 28B(5):355–364, 1994.

Gruber, T. R.: Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human Computer Studies, 43(5):907–928, 1995.

Han, J., Cheng, H., Xin, D., and Yan, X.: Frequent pattern mining: current status and future directions. Data Mining and Knowledge Discovery, 15(1):55–86, August 2007.

Harms, S. K., Deogun, J., and Tadesse, T.: Discovering Sequential Association Rules with Constraints and Time Lags in Multiple Sequences. In 13th International Symposium on Foundations of Intelligent Systems: ISMIS 2002, volume 2366/2002, pages 373–376. Lyon, France, Springer Berlin / Heidelberg, 2002.

Harms, S. K. and Deogun, J. S.: Sequential Association Rule Mining with Time Lags. Journal of Intellegent Information Systems, 22(1):7–22, 2004.

eds. D. Hensher and P. Stopher Behavioural travel modelling. Taylor & Francis, 1979.

Hipp, J., Güntzer, U., and Nakhaeizadeh, G.: Algorithms for association rule mining a general survey and comparison. SIGKDD Explorations Newsletter, 2(1):58–64, 2000.

Hotho, A., Nürnberger, A., and Paaß, G.: A Brief Survey of Text Mining. GLDV-Journal for Computational Linguistics and Language Technologie, 20:19–62, 2005.

Hovy, E.: The Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses. In Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC), 1998.

Hu, W. and Qu, Y.: Discovering Simple Mappings Between Relational Database Schemas and Ontologies. In The Semantic Web, volume 4825/2008 of Lecture Notes in Computer Science, pages 225–238. Springer Berlin / Heidelberg, October 2007.

iSuppli: Shipments of GPS-Enabled Mobile Handsets to More than Quadruple by 2011. http://www.isuppli.com/MarketWatch/Pages/Shipments-of-GPS-Enabled-Mobile-Handsets-to-More-than-Quadruple-by-2011.aspx?id=63, November 2007.

Janssens, D., Wets, G., Brijs, T., Vanhoof, K., Arentze, T., and Timmermans: Improving the Performance Of Multiagent Rule-Based Model For Activity Pattern Decisions With Bayesian Networks. Transportation Research Record, Journal of the Transportation Research Board, 1894:75–83, 2004.

Janssens, D., Lan, Y., Wets, G., and Chen, G.: The optimization of activity-travel sequences by means of reinforcement learning. In International Conference on Intelligent Systems and Knowledge Engineering, Shangai, China, 2006.

Janssens, D., Lan, Y., Wets, G., and Chen, G.: Allocating time and location information to activity-travel patterns through reinforcement learning. Knowledge Based Systems, 20(5):466–477, 2007.

Janssens, D., Wets, G., Brijs, T., and Vanhoof, K.: Simulating Daily Activity Patterns Through the Identification of Sequential Dependencies. In Progress in Activity-Based Analysis, ed. H. Timmermans, pages 67–89. Elsevier, 2005.

Jindal, N. and Liu, B.: Mining Comparative Sentences and Relations. In Proceedings of the Twenty-First National Conference on Artificial Intelligence, 2006.

Joh, C., Arentze, T., Hofman, F., and Timmermans, H.: Activity pattern similarity: a multidimensional sequence alignment method. Transportation Research Part B, 36(5):385–403, 2002.

Joh, C., Arentze, T., and Timmermans, H.: Multidimensional sequence alignment methods for activity-travel pattern analysis: A comparison of dynamic programming and genetic algorithms. Geographical Analysis, 33:247–270, 2001.

Joh, C., Arentze, T., and Timmermans, H.: A utility-based analysis of activity time allocation decisions underlying segmented daily activity-travel patterns. A utility-based analysis of activity time allocation decisions underlying segmented daily activity-travel patterns., 37:105–126, 2005.

Joh, C.-H., Arentze, T., and Timmermans, H.: Pattern Recognition in Complex Activity Travel Patterns: Comparison of Euclidean Distance, Signal-Processing Theoretical, and Multidimensional Sequence Alignment Methods. Transportation Research Record: Journal of the Transportation Research Board, 1752:16–22, January 2001.

Joh, C.-H., Arentze, T., and Timmermans, H.: Understanding activity scheduling and rescheduling behaviour: Theory and numerical illustration. GeoJournal, 53(4):359–371, April 2001.

Jones, P., Koppelman, F., and Orfeuil, J.: Activity analysis: State-of-the-art and future directions. In Developments in Dynamic and Activity-Based Approaches to Travel, ed. P. Jones. Avebury, 1990.

Kawaguchi, N.: WiFi Location Information System for Both Indoors and Outdoors. Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living, 5518:638–645, 2009.

Keuleers, B., Wets, G., Arentze, T., and Timmermans, H.: Association rules in identification of spatial-temporal patterns in multiday activity diary data. Transportation Research Record: Journal of the Transportation Research Board, 1752:32–37, 2001.

Kitamura, R., Chen, C., and Pendyala, R.: Generation of Synthetic Daily Activity-Travel Patterns. Transportation Research Record, 1607:154–162, 1997.

Kitamura, R., Yamamoto, T., Susilo, Y., and Axhausen, K.: How routine is a routine? An analysis of the day-to-day variability in prism vertex location. Transportation Research Part A, 40(3):259–279, 2006.

Kitamura, R.: An evaluation of activity-based travel analysis. Transportation, 15(1):9–34, March 1988.

Kwan, M. P. and Golledge, R.: Integration of GIS with Activity-Based Model in Atis. In 74th Annual Meeting of Transportation Research Board, Washington, DC, 1995.

Lahiri, M. and Berger-Wolf, T. Y.: Periodic subgraph mining in dynamic networks. Knowledge and Information Systems, September 2009.

Lakshminarayan, K., Harp, S., Goldman, R., Samad, T., et al.: Imputation of missing data using machine learning techniques. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pages 140–145, 1996.

Lee, M. S. and McNally, M. G.: On the structure of weekly activity/travel patterns. In Transportation Research Part A: Policy and Practice, volume 37, pages 823–839, 2003.

Lee, M. and McNally, M.: An empirical investigation on the dynamic processes of activity scheduling and trip chaining. Transportation, 33(6):553–565, November 2006.

Lee, M. and McNally, M.: Experiments with computerized self-administrative activity survey. Transportation Research Record: Journal of the Transportation Research Board, 1752:91–99, 2001.

Li, X.-L. and Liu, B.: Machine Learning: ECML 2005, volume 3720 of Lecture Notes in Computer Science, chapter Learning from Positive and Unlabeled Examples with Different Data Distributions, pages 218 – 229. Springer Berlin / Heidelberg, 2005.

Liao, L.: Location-based Activity Recognition. Doctoral dissertation, University of Washington, 2006.

Liao, L., Fox, D., and Kautz, H.: Learning and Inferring Transportation Routines. In Proceedings of the Nineteenth National Conference on Artificial Intelligence, pages 348–353, San Jose, California, July 2004.

Liao, L., Patterson, D. J., Fox, D., and Kautz, H.: Learning and inferring transportation routines. Artificial Intelligence, 171(5-6):311–331, April 2007.

Liu, B.: Web data mining: exploring hyperlinks, contents, and usage data. Data-Centric Systems and Applications. Springer, 2007.

Liu, B., Hsu, W., and Ma, Y.: Mining association rules with multiple minimum supports. In KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 337–341, New York, NY, USA, 1999. ACM Press.

Liu, B., Hu, M., and Cheng, J.: Opinion observer: analyzing and comparing opinions on the Web. In WWW '05: Proceedings of the 14th international conference on World Wide Web, pages 342–351, New York, NY, USA, 2005. ACM Press.

Liu, G. Y. and Maguire, G. Q.: Efficient Mobility Management Support for Wireless Data Services. In Proc. of 45th IEEE Vehicular Technology Conference, Chicago, Illinois, 1995.

Liu, G. and Maguire, G.Q., J.: A predictive mobility management algorithm for wireless mobile computing and communications. In Universal Personal Communications. 1995. Record., 1995 Fourth IEEE International Conference on, pages 268–272, 1995.

Macker, J. P. and Corson, M. S.: Mobile ad hoc networking and the IETF. SIGMOBILE Mob. Comput. Commun. Rev., 2(1):9–14, 1998.

Madre, J.-L., K. A. and Brg, W.: Immobility in travel diary surveys. Transportation, Vol.34:107–128, 2007.

Marca, J. E., Rindt, C. R., and McNally, M. G.: Collecting Activity Data from GPS Readings. Technical Report Paper UCI-ITS-AS-WP-02-3, Institute of Transportation Studies, Center for Activity Systems Analysis, University of California, Irvine, July 2002.

Marshall, B., Chen, H., and Madhusudan, T.: Matching knowledge elements in concept maps using a similarity flooding algorithm. Decision Support Systems, 42(3):1290–1306, December 2006.

McNally, M. G.: An Activity-Based Microsimulation Model for Travel Demand Forecasting. Technical report, UC Irvine: Center for Activity Systems Analysis, May 1996.

McNally, M. G.: The Activity-Based Approach. Technical Report Paper UCI-ITS-AS-WP-00-4, Center for Activity Systems Analysis, University of California, Irvine, December 2000.

Meister, K., Frick, M., and Axhausen, K. W.: A GA-based household scheduler. Transportation, 32(5):473–494, 2005.

Melnik, S., Garcia-Molina, H., and Rahm, E.: Similarity flooding: a versatile graph matching algorithm and its application to schema matching. In Data Engineering, 2002. Proceedings. 18th International Conference on, pages 117–128, 2002.

Miller, E. and Roorda, M.: Prototype Model of Household Activity-Travel Scheduling. Transportation Research Record, 1831:114–121, 2003.

Mitchell, R. B. and Rapkin, C.: Urban traffic: A function of land used. Columbia University Press, 1954.

Mobasher, B., Dai, H., Luo, T., and Nakagawa, M.: Using Sequential and Non-Sequential Patterns in Predictive Web Usage Mining Tasks. In ICDM '02: Proceedings of the 2002 IEEE International Conference on Data Mining, page 669, Washington, DC, USA, 2002. IEEE Computer Society.

Mohammadian, A. and Doherty, S. T.: A Mixed Logit Model of Activity Scheduling Time Horizon incorporating spatial-temporal variables. Journal of the Transportation Research Board, pages 33–40, 2005.

Mohammadian, A. and Zhang, Y.: Investigating the Transferability of National Household Travel Survey Data. Transportation Research Record: Journal of the Transportation Research Board, 1993:67–79, 2007.

Mohammadian, A. and Zhang, Y.: Transferability of National Household Travel Survey Data to Local Areas. In Proc. 85th Annual Meeting of the Transportation Research Board, Washington, D.C., January 2006.

Mohammadian, A. and Doherty, S. T.: Modeling activity scheduling time horizon: Duration of time between planning and execution of pre-planned activities. Transportation Research Part A: Policy and Practice, 40(6):475–490, July 2006.

Mohammadian, K., Rashidi, T., and Takuriah, P.: Effectiveness of transit strategies targeting elderly people: Survey results and preliminary data analysis. Technical Report FHWA-ICT-09-033 a report of the findings of the project ICT-R27-17, Illinois Center for Transportation, February 2009.

Murakami, E. and Wagner, D.: Can using global positioning system (GPS) improve trip reporting? Transportation Research Part C, 7(2-3):149–165, 1999.

Nijland, E. W. L., Arentze, T. A., Borgers, A. W. J., and Timmermans, H. J. P.: Individuals' activity - travel rescheduling behaviour: experiment and model-based analysis. Environment and Planning A, 41(6):1511–1522, 2009.

North, R., Richards, M., Cohen, J., Hoose, N., Hassard, J., and Polak, J.: A mobile environmental sensing system to manage transportation and urban air quality. Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on, pages 1994–1997, May 2008.

Noy, N. F.: Semantic integration: a survey of ontology-based approaches. SIGMOD Rec., 33(4):65–70, 2004.

NuStats: 2002 Anchorage Household Travel Survey Technical Report. Technical report, Municipality of Anchorage, September 2002.

NuStats: 2001 Atlanta Household Travel Survey: Final Report. Technical report, Atlanta Regional Commision, April 2003.

Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., and Hsu, M.-C.: Mining sequential patterns by pattern-growth: the PrefixSpan approach. Knowledge and Data Engineering, IEEE Transactions on, 16:1424–1440, 2004.

Pei, J., Han, J., and Wang, W.: Mining sequential patterns with constraints in large databases. In CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management, pages 18–25, New York, NY, USA, 2002. ACM.

Pendyala, R. M. and Bhat, C. R.: An Exploration of the Relationship between Timing and Duration of Maintenance Activities. Transportation, 31(4):429–456, 2004.

Pribyl, O. and Goulias, K. G.: Simulation of Daily Activity Patterns Incorporating Interactions Within Households: Algorithm Overview and Performance. In Transportation Research Record, volume 1926, pages 135–141, January 2005.

Quddus, M. A., Ochieng, W. Y., Zhao, L., and Noland, R. B.: A general map matching algorithm for transport telematics applications. GPS Solutions, 7(3):157 – 167, 2003.

Ragel, A. and Crmilleux, B.: MVC–a preprocessing method to deal with missing values. Knowledge-Based Systems, 12(5-6):285 – 291, 1999.

Ragel, A. and Crmilleux, B.: Treatment of missing values for association rules. Research and Development in Knowledge Discovery and Data Mining, 1394:258–270, 1998.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J., and Solenberger, P.: A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey methodology, 27(1):85–96, 2001.

Rahm, E. and Bernstein, P. A.: A survey of approaches to automatic schema matching. The VLDB Journal, 10(4):334–350, 2001.

Recker, W., McNally, M., and Root, G.: A Model of complex travel behavior. Transportation Research A, 20(4), 1986.

Research, A.: GPS-Enabled Handsets Expected to Bypass the Economic Downturn. http://www.abiresearch.com/press/1351-GPS-enabled+Handsets+Expected+to+Bypass+the+Economic+Downturn, January 2009.

Richardson, A. J., Ampt, E. S., and Meyburg, A. H.: Nonresponse issues in household travel surveys. In Conference on Household Travel Surveys: New Concepts and Research Needs, pages 79–114. Transportation Research Board, January 1996.

Rubin, D.: Inference and missing data. Biometrika, 63(3):581, 1976.

Rudloff, C. and Ray, M.: Detecting Travel Modes and Profiling Commuter Routes Solely Based on GPS Data. In Proceedings of 89th Annual Meeting of the Transportation Research Board, 2010.

Ruiz, T. and Timmermans, H.: Changing the timing of activities in resolving Scheduling Conflicts. Transportation, 33(5):429–445, 2006.

Schafer, J. L. and Graham, J. W.: Missing data: Our view of the state of the art. Psychological Methods, 7(2):147–177, June 2002.

Schönfelder, S., Li, H., Guensler, R., Ogle, J., and Axhausen, K.: Analysis of commute Atlanta instrumented vehicle GPS data: Destination choice behavior and activity spaces. In 85th Annual Meeting of the Transportation Research Board, Washington, DC, 2006.

Schuessler, N. and Axhausen, K.: Processing Raw Data from Global Positioning Systems Without Additional Information. Transportation Research Record: Journal of the Transportation Research Board, 2105:28–36, December 2009.

Shen, J.-J., Chang, C.-C., and Li, Y.-C.: Combined association rules for dealing with missing values. Journal of Information Science, 33(4):468–480, 2007.

Shoval, N. and Isaacson, M.: Sequence Alignment as a Method for Human Activity Analysis in Space and Time. Annals of the Association of American Geographers, 97(2):282–297, 2007.

Shoval, N. and Raveh, A.: The categorization of tourist attractions: The modeling of tourist cities based on a new method of multivariate analysis. Tourism Management, 25(6):74150, 2004.

Shvaiko, P. and Euzenat, J.: A Survey of Schema-based Matching Approaches. Journal on Data Semantics (JoDS), IV(LNCS 3730):146–171, 2005.

Song, C., Qu, Z., Blumm, N., and Barabsi, A.-L.: Limits of Predictability in Human Mobility. Science, 327(5968):1018–1021, February 2010.

Srikant, R. and Agrawal, R.: Mining Sequential Patterns: Generalizations and Performance Improvements. In Proc. 5th Int. Conf. Extending Database Technology, EDBT, eds. P. M. G. Apers, M. Bouzeghoub, and G. Gardarin, volume 1057, pages 3–17. Springer-Verlag, 25–29 1996.

Stoilos, G., Stamou, G., and Kollias, S.: A String Metric for Ontology Alignment. In The Semantic Web ISWC 2005, volume 3729/2005 of Lecture Notes in Computer Science, pages 624 – 637. Springer Berlin / Heidelberg, October 2005.

Stopher, P., Jiang, Q., and FitzGerald, C.: Processing GPS Data from Travel Surveys. In 28th Australasian Transport Research Forum, Sydney, September 2005.

Stopher, P. R., Jiang, Q., and FitzGerald, C.: Processing GPS data from travel surveys. In paper presented at 2nd International Colloqium on the Behavioural Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications, Toronto, June 2005.

Stopher, P.: Use of an activity-based diary to collect household travel data. Transportation, 19(2):159–176, May 1992.

Stopher, P., FitzGerald, C., and Xu, M.: Assessing the accuracy of the Sydney Household Travel Survey with GPS. Transportation, 34(6):723–741, 11 2007.

Stopher, P., Kockelman, K., Greaves, S., and Clifford, E.: Reducing Burden and Sample Sizes in Multiday Household Travel Surveys. Transportation Research Record: Journal of the Transportation Research Board, 2064(1):12–18, December 2008.

Stopher, P. R., Alsnih, R., Wilmot, C. G., Stecher, C., Pratt, J., Zmud, J., Mix, W., Freedman, M., Axhausen, K., Lee-Gosselin, M., Pisarski, A. E., and Brg, W.: Standardized procedures for personal travel surveys. Technical Report National Cooperative Highway Research Program report 571, Transportation Research Board of the National Academies, Washington, D.C., 2008.

Sunna, W. and Cruz, I. F.: Using the AgreementMaker to Align Ontologies for the OAEI Campaign 2007. In Proceedings of the 2nd International Workshop on Ontology Matching (OM-2007), Busan, Korea, November 2007.

Susilo, Y. and Kitamura, R.: Analysis of Day-to-Day Variability in an Individual's Action Space: Exploration of 6-Week Mobidrive Travel Diary Data. Transportation Research Record: Journal of the Transportation Research Board, 1902:124–133, 2005.

Timmermans, H., Waerden, P. V. D., Alves, M., Polak, J., Ellis, S., Harvey, A., Kurose, S., and Zandee, R.: Spatial Context and the Complexity of Daily Travel Patterns: An International Comparison. Journal of Transport Geography, 11(1):37–46, 2003.

ed. H. Timmermans Progress in Activity-Based Analysis. Oxford,England, Elsevier, 2005.

Torrens, M., Hertzog, P., Pu, P., and Faltings, B.: Towards an intelligent mobile travel assistant. In SAC '04: Proceedings of the 2004 ACM symposium on Applied computing, pages 1208–1209, New York, NY, USA, 2004. ACM Press.

Tresp, V., Bundschus, M., Rettinger, A., and Huang, Y.: Towards Machine Learning on the Semantic Web. In Uncertainty Reasoning for the Semantic Web I, volume 5327 of Lecture Notes in Computer Science, pages 282–314. Springer Berlin / Heidelberg, 2008.

Tsui, S. Y. A. and Shalaby, A. S.: Enhanced System for Link and Mode Identification for Personal Travel Surveys Based on Global Positioning Systems. Transportation Research Record: Journal of the Transportation Research Board, 1972:38–45, 2006.

van Rijsbergen, C.: Information Retrieval. London, Butterworth, 1979.

Varshavsky, A., Chen, M., de Lara, E., Froehlich, J., Haehnel, D., Hightower, J., LaMarca, A., Potter, F., Sohn, T., Tang, K., and Smith, I.: Are GSM Phones THE Solution for Localization? In WMCSA '06: Proceedings of the Seventh IEEE Workshop on Mobile Computing Systems and Applications, pages 34–42, Washington, DC, USA, 2006. IEEE Computer Society.

Vovsha, P., Bradley, M., and Bowman, J.: Activity-based travel forecasting models in the United States: Progress since 1995 and Prospects for the Future. In EIRASS Conference on Progress in Activity-Based Analysis, Maastricht, The Netherlands, 2004.

Wang, D. and Cheng, T.: A spatio-temporal data model for activity-based transport demand modelling. International Journal of Geographical Information Science, 15(6):561–585, September 2001.

Wen, C.-H. and Koppelman, F.: A conceptual and methdological framework for the generation of activity-travel patterns. Transportation, 27(1):523, February 2000.

Williams, C. A., Mohammadian, A., Nelson, P. C., and Doherty, S. T.: Mining Sequential Association Rules for Traveler Context Prediction. In Proceedings of the First International Workshop on Computational Transportation Science held at The International Conference on Mobile and Ubiquitous Systems: Networks and Services (MOBIQUITOUS 2008), Dublin, Ireland, July 2008.

Williams, C. A., Nelson, P. C., and Mohammadian, A.: Attribute Constrained Rules for Partially Labeled Sequence Completion. Advances in Data Mining - Applications and Theoretical Aspects, 5633:338 – 352, July 2009.

Wolf, J.: Using GPS Data Loggers to Replace Travel Diaries in the Collection of Travel Data. Doctoral dissertation, Georgia Institute of Technology, School of Civil and Environmental Engineering, Atlanta, Georgia, July 2000.

Wolf, J., Guensler, R., and Bachman, W.: Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data. Transportation Research Record: Journal of the Transportation Research Board, 1768:125–134, 2001.

Yang, Q., Zhang, H. H., and Li, T.: Mining web logs for prediction models in WWW caching and prefetching. In KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pages 473–478, New York, NY, USA, 2001. ACM.

VITA

Chad A. Williams

chadwilliams13@gmail.com

EDUCATION

Ph.D. University of Illinois at Chicago 2006 – 2010
Computer Science
Dissertation: A Data Mining Approach To Rapidly Learning Traveler Activity Patterns For Mobile Applications
Advisors: Peter Nelson and Abolfazl (Kouros) Mohammadian

M.S. DePaul University 2004 – 2006
Computer Science (with distinction)
Thesis: Profile Injection Attack Detection for Securing Collaborative Recommender Systems
Advisor: Bamshad Mobasher

B.S. Cornell University 1994 - 1998
Computer Science

FELLOWSHIPS/AWARDS/HONORS

NSF IGERT Fellow 2006 – 2010
Integrative Graduate Education and Research Traineeship (IGERT), Interdisciplinary education spanning science, technology, engineering, mathematics and social sciences

Best Paper Award, The 8th IEEE Conference on E-Commerce Technology (CEC) 2006

Best Paper Award, DePaul CTI Research Symposium / Midwest Software Engineering Conference 2006

Phi Kappa Phi Honor Society

Upsilon Pi Epsilon, International Honor Society for the Computing and Information Disciplines

RESEARCH EXPERIENCE

NSF IGERT Research Fellow, University of Illinois at Chicago 2006 - 2010
Initiated interdisciplinary effort aimed at learning an individual travelers activity patterns culminating in an international collaboration across multiple disciplines including computer science, transportation planning and geographic analysis. Managed undergraduate and graduate students in research effort.

Research Assistant, DePaul University 2004 - 2006
Participated in drafting a NSF project proposal and explored the vulnerabilities of recommender systems introducing techniques for recognizing attacks and minimizing their impact

Research Programmer, Cornell University 1997 - 1998
Developed distributed version control system for the Ensemble project similar to CVS in order to demonstrate potential applications of a secure group communication protocol

TEACHING EXPERIENCE

Teaching Associate, University of Illinois at Chicago 2006
Introduction to Software Engineering, Co-instructed class of 20 students, preparing and presenting 60% of the lectures. Developed the syllabus, managed group projects, wrote and graded the final exam and directed work of teaching assistant.

Instructor, BlueMeteor, Inc. 2000 - 2001
New developer training, Instructed, developed training material and practice projects for 6 employees

Development Lead, Andersen Consulting (n/k/a Accenture) 2000 - 2001
New developer training, Instructed and developed training material for 20 employees joining the team

JOURNAL PUBLICATIONS

1. M. Z. Frignani, J. Auld, A. Mohammadian, C. Williams and P. C. Nelson. Competing Hazard Model of Household Vehicle Transaction Behavior with Discrete Time Intervals and Unobserved Heterogeneity, accepted for publication in Transportation Research Record, 2010.
2. C. A. Williams, P. C. Nelson and A. Mohammadian. Attribute Constrained Rules for Partially Labeled Sequence Completion, Advances in Data Mining - Applications and Theoretical Aspects, vol. 5633 of Lecture Notes in Computer Science, (Petra Perner, ed.), July 2009, pp. 338 - 352.
3. J. Auld, C. A. Williams, A. Mohammadian and P. C. Nelson. An Automated GPS-Based Prompted Recall Survey With Learning Algorithms, Transportation Letters: The International Journal of Transportation Research, vol. 1, no. 1, Jan. 2009, pp. 59-79.
4. C. Williams, B. Mobasher and R. Burke. Defending Recommender Systems: Detection of Profile Injection Attacks, Service Oriented Computing and Applications, vol. 1, no. 3, Nov. 2007, pp. 157-170.
5. B. Mobasher, R. Burke, R. Bhaumik and C. Williams. Toward Trustworthy Recommender Systems: An Analysis of Attack Models and Algorithm Robustness, ACM Transactions on Internet Technology, vol. 7, no. 4, Oct. 2007, ACM.
6. B. Mobasher, R. Burke, C. Williams and R. Bhaumik. Analysis and Detection of Segment-Focused Attacks Against Collaborative Recommendation, Advances in Web Mining and Web Usage Analysis, vol. 4198 of Lecture Notes in Artificial Intelligence, (O. R. Zaane, O. Nasraoui and P. S. Yu, eds.), 2006, pp. 96-118.

REFEREED CONFERENCE PUBLICATIONS

1. M. Z. Frignani, J. Auld, A. Mohammadian, C. Williams and P. C. Nelson. Urban Travel Route and Activity Choice Survey (UTRACS): An Internet-Based Prompted Recall Activity Travel Survey using GPS Data, Proceedings of the 89th Annual Meeting of the Transportation Research Board, (DVD), Washington, D.C., Jan. 2010.
2. C. A. Williams, A. Mohammadian, P. C. Nelson and S. T. Doherty. Mining Sequential Association Rules for Traveler Context Prediction, Proceedings of the First International Workshop on Computational Transportation Science, Held at The International Conference on Mobile and Ubiquitous Systems: Networks and Services (MOBIQUITOUS 2008), Dublin, Ireland, July 2008.
3. R. Burke, B. Mobasher, C. Williams and R. Bhaumik. Classification Features for Attack Detection in Collaborative Recommender Systems, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, Pennsylvania, 2006, pp. 542-547.
4. C. Williams, R. Bhaumik, R. Burke and B. Mobasher. The Impact of Attack Profile Classification on the Robustness of Collaborative Recommendation, Proceedings of the 2006 WebKDD Workshop, Held at KDD 2006, Philadelphia, Pennsylvania, Aug. 2006.
5. C. Williams, B. Mobasher, R. Burke, J. Sandvig and R. Bhaumik. Detection of Obfuscated Attacks in Collaborative Recommender Systems, Proceedings of the ECAI06 Workshop on Recommender Systems, Held at the 17th European Conference on Artificial Intelligence (ECAI'06), Riva del Garda, Italy, Aug. 2006.
6. R. Bhaumik, C. Williams, B. Mobasher and R. Burke. Securing Collaborative Filtering Against Malicious Attacks Through Anomaly Detection, Proceedings of the 4th Workshop on Intelligent Techniques for Web Personalization (ITWP'06), Held at AAAI 2006, Boston, Massachusetts, July 2006.
7. R. Burke, B. Mobasher, C. Williams and R. Bhaumik. Detecting Profile Injection Attacks in Collaborative Recommender Systems, Proceedings of the 8th IEEE Conference on E-Commerce Technology (CEC'06), San Francisco, California, June 2006.
   ** Winner of Best Paper Award*
8. C. Williams, R. Bhaumik, J. Sandvig, B. Mobasher and R. Burke. Evaluation of Profile Injection Attacks in Collaborative Recommender Systems, DePaul CTI Research Symposium /Midwest Software Engineering Conference (CTIRS/MSEC 2006), Chicago, Illinois, Apr. 2006.
   ** Winner of Best Paper Award*
9. R. Burke, B. Mobasher, R. Bhaumik and C. Williams. Segment-Based Injection Attacks Against Collaborative Filtering Recommender Systems, Proceedings of the 2005 International Conference on Data Mining (ICDM'05), Houston, Texas, Nov. 2005.
10. R. Burke, B. Mobasher, R. Bhaumik and C. Williams. Collaborative Recommendation Vulnerability to Focused Bias Injection Attacks, Proceedings of the Workshop on Privacy and Security Aspects of Data Mining, Held at ICDM'05, Houston, Texas, Nov. 2005.

11. B. Mobasher, R. Burke, R. Bhaumik and C. Williams. Effective Attack Models for Shilling Item-Based Collaborative Filtering Systems, Proceedings of the 2005 WebKDD Workshop, Held at KDD 2005, Chicago, Illinois, Aug. 2005.

CONFERENCE PRESENTATIONS AND INVITED TALKS

1. A Framework for Traveler Pattern Prediction, invited talk to be presented at Computation Transportation Science seminar at Schloss Dagstuhl - Leibniz Center for Informatics, Dagstuhl Germany, March 2010.
2. Learning Activity Patterns of Individuals, presented at Doctoral Student Research in Transportation Modeling at Transportation Research Board 89th Annual Meeting, January 2010.
3. Attribute Constrained Rules for Partially Labeled Sequence Completion, paper presented at Industrial Conference on Data Mining, Leipzig, Germany, July 2009.
4. Attribute Constrained Rules: A New Approach for Missing Traveler Data, talk presented at University of Illinois at Chicago, Department of Computer Science Colloquium, July 2009.
5. Learning Travel Patterns of Individuals, talk presented at University of Illinois at Chicago, IGERT Seminar Series, January 2009.
6. Mining Sequential Association Rules for Traveler Context Prediction, paper presented at First International Workshop on Computational Transportation Science, Dublin, Ireland, July 2008.
7. Quickly Learning Activity and Travel Patterns of Individuals: Transfer Learning for Individual Travel Behavior Prediction, talk presented at University of Illinois at Chicago, IGERT Seminar Series, February 2008.
8. Effective Attack Models for Shilling Item-Based Collaborative Filtering Systems, paper presented at WebKDD Workshop held at KDD 2005, Chicago, Illinois, August 2005.

OTHER PUBLICATIONS AND POSTERS

1. Computational Transportation Science: An Interdisciplinary Approach to Integrating Emerging Technologies into Transportation, by Chad A. Williams, Ouri Wolfson and Peter C. Nelson, Poster presented at 2008 NSF IGERT Project Meeting, Arlington, Virgina, May 2008.
2. Genetically Evolving Optimal Neural Networks, by Chad Williams. In Neural Networks and Expert Systems, The Institute of Chartered Financial Analysts of India (ICFAI), Jan. 2007.

RELATED WORK EXPERIENCE

Manager, Accenture 2001 - 2004

Worked in India for 3 months as team lead in Accentures initial offshore development effort. Created interdisciplinary computational solutions for Fortune 500 companies in insurance, capital markets, banking and credit reporting. Developed project plans and led teams as large as 20 developers through project lifecycle to delivery.

Technical Architect, BlueMeteor, Inc. 2000 - 2001

Helped write proposals that resulted in selling projects to three Fortune 500 companies

Consultant, Andersen Consulting (n/k/a Accenture) 1998 - 2000

Developed application and database design for multiple Fortune 500 insurance companies

PROFESSIONAL ACTIVITIES

President, DePaul Student Chapter of IEEE 2004 - 2006

Member, IEEE

Computational Intelligence Society, Intelligent Transportation Systems Society

Member, Association for Computing Machinery (ACM)

SIGART, SIGKDD, SIGMOBILE

Member, American Association for Artificial Intelligence (AAAI)