

DEVELOPING A TRIP DISTRIBUTION MODEL FOR IDENTIFIED MOBILITY GROUPS USING BIG DATA

Buddhi Ayesha Rathnayaka

188006L

Thesis/Dissertation submitted in partial fulfillment of the requirements for the
Degree of Master of Science in Computer Science and Engineering

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

February 2020

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date:

The above candidate has carried out research for the Masters thesis/dissertation under my supervision.

Name of the Supervisor: Dr. Charith Chitraranjan

Signature of the Supervisor:

Date:

Name of the Supervisor: Dr. Amal Shehan Perera

Signature of the Supervisor:

Date:

Name of the Supervisor: Prof. Amal S. Kumarage

Signature of the Supervisor:

Date:

ACKNOWLEDGEMENT

First and foremost, I would like to thank my supervisors Dr. Amal Shehan Perera, Dr. Charith Chitraranjan and Prof. Amal S. Kumarage for the continuous support and guidance I received in every aspect while completing this research. I could not have imagined having a better set of advisors for my research. I am highly grateful for their insightful guidance and encouragement provided throughout my research work.

I would also like to thank my progress review committee, Dr. Lochandaka Ranatunge, and Dr. Dilum Bandara for their valuable insights and guidance. Their advice helped me to improve the state of my research work.

A very special gratitude goes to Sriganesh Lokanathan (Team leader, LIRNEasia), and I also thank Lasantha Fernando (Researchers - LIRNEasia). Further, I would like to extend my gratitude towards LIRNEasia organizations for all the support given during the study period.

This work was funded by a Senate Research Committee (SRC) Grant of the University of Moratuwa. Further, I would like to thank LK Domain Registry for the support given for the conference participation via travel grants.

I am sincerely thankful to all the staff members and the non-academic staff of the Department of Computer Science and Engineering. Finally, I would like to thank my fellow lab mates in Data Science and Engineering Research Hub of University of Moratuwa and all of my friends and family. Not just academic work, I was able to enjoy my time while involving different activities and outings. It helped me to keep my life balanced. I thank my parents who have given me a very fortunate life and always believing, trusting and supporting me.

ABSTRACT

The need for frequent transportation planning has become a key factor since people started becoming more mobile making urban traffic patterns more complex. The primary source for analysing such travel behavior is through manual surveys. These surveys are expensive, time consuming and often are outdated by the time the survey is completed for analysis. To overcome these issues, Mobile Network Big Data (MNBD) which concerns large data sets can be used over such traditional data collection processes. Call Detail Records (CDR) which is a subset of MNBD is readily available as most of the telecommunication service providers maintain CDR. Thus, analyzing CDR leads to an efficient identification of human behavior and location.

However, many researches on CDRs have been done focusing to identify travel patterns in order to understand human mobility behavior. Relatively high percentage of sparse data and other scenarios like the Load Sharing Effect (LSE) causes difficulties in identifying precise location of the user when using CDR data. Existing approaches for identifying precise user location patterns have certain constraints. Past researches utilizing CDRs have used primary approaches in recognizing load sharing effects and have given minimum consideration to the transmission power of the respective cell towers when localizing the users. Furthermore, these studies have neglected the differences in mobility behavior of different segment of users and taken the entire community of users as a single cluster.

In this research, a novel methodology to overcome these limitations is introduced for locating users from CDRs by dividing the users into distinct clusters for identifying the model parameters and through enhanced identification of load sharing effects by taking the transmission power into consideration. Further, this study contributes to the transport sector by identifying secondary activities from CDR data, without limiting to the primary activity recognition. This research uses approximately 4 billion CDR data points, voluntarily collected mobile data and manually collected travel survey data to find techniques to overcome the existing limitations and validate the results.

Proposed dynamic filtering algorithm for load shared records identification showed a significant improvement on accuracy over previous predefined speed based filtering methods. Further, we found that, IO-HMM outperforms standard HMM results on activity recognition.

Keywords: Travel Demand Modeling with Mobile Network Big Data, User Localization Based on CDRs, Activities identification from CDRs.

LIST OF FIGURES

Figure 1.1	Four-step model	3
Figure 1.2	Characteristics of residential and non-residential zones	4
Figure 3.1	Flow diagram for the overall methodology	21
Figure 3.2	CDR mobile app listed on the Google play store	25
Figure 3.3	CDR mobile app supporting three languages	25
Figure 3.4	Data collection at the user registration interface	25
Figure 3.5	Pop-up notifications during the initiation of travel	25
Figure 3.6	Data collection with the initiation of travel	26
Figure 3.7	Permission request to use the GPS navigation	26
Figure 3.8	Load sharing effect	31
Figure 3.9	Clustering-based approach for western province	33
Figure 3.10	Use locations based on the number of appearances days	36
Figure 3.11	Use locations based on the signal strength	36
Figure 3.12	Stay locations based on CDR and GPS	40
Figure 3.13	Aggregated accuracy with caller activity levels for weekdays	42
Figure 3.14	Aggregated accuracy with caller activity levels for weekends	42
Figure 3.15	Travelling and non-travelling hour prediction accuracy with average frequency of CDR for weekdays	43
Figure 3.16	Travelling and non-travelling hour prediction accuracy with average frequency of CDR for weekends	46
Figure 3.17	Accuracy of travelling and non-travelling hours	46
Figure 3.18	Average travel distance based on HVS data	47
Figure 3.19	Average travel distance based on CDR data	47
Figure 3.20	Activity generation time-window analysis for CDR and HVS data	48
Figure 3.21	IO-HMM architecture	48
Figure 3.22	Estimation step	50
Figure 3.23	Maximization step	50
Figure 3.24	Parameters update step	50

Figure 4.1	Speed vs Error after removing load sharing effect for CDR dataset	52
Figure 4.2	Home localization error	53
Figure 4.3	Work localization error	54

LIST OF TABLES

Table 1.1	Types of trip movements	5
Table 1.2	CDR data parameters	6
Table 2.1	Identifying significant locations using CDR in transport aspects	15
Table 2.2	O-D matrix estimations using CDR in transport aspects	16
Table 3.1	Data collected at the user registration interface	23
Table 3.2	Data collected with the initiation of travel	24
Table 3.3	Automatically collected data through background processing	26
Table 3.4	Data collected through the mobile app	27
Table 3.5	Description of trip purposes	30
Table 3.6	Distribution of home-based work trips	30
Table 3.7	Methodology to minimize the Load sharing effect	32
Table 3.8	Accuracy of stay locations identified from CDR	41
Table 3.9	Analysis of travelling and non-travelling hours for weekdays	44
Table 3.10	Analysis of travelling and non-travelling hours for weekends	45
Table 4.1	User profiling results comparison (Precision, Recall and F1 Score)	51
Table 4.2	Load sharing record identification results comparison with mobile app data	52
Table 4.3	Home work distribution - HVS data	55
Table 4.4	Home work distribution - Call days based method	55
Table 4.5	Home work distribution - Proposed methodology	55
Table 4.6	Activity recognition models accuracy	56
Table 4.7	Primary and Secondary activities recognition accuracy	56

LIST OF ABBREVIATIONS

CA	Call Activity
CAC	Call Activity Count
CDR	Call Detail Records
CLS	Cordon Line Surveys
DSD	Divisional Secretariat Divisions
HBW	Home-based Work
HMM	Hidden Markov Model
HVS	Household Visit Survey
IO-HMM	Input/Output Hidden Markov Model
JICA	Japan International Cooperation Agency
LSE	Load Sharing Effect
LSR	Load Sharing Records
MNBD	Mobile Network Big Data
O-D	Origin-Destination
PCA	Principal Component Analysis
SLS	Screen Line Survey
SVM	Support Vector Machine
TGS	Trip Generation Survey
TSS	Travel Speed Survey
VLR	Visitor Location Register

TABLE OF CONTENTS

Declaration of the Candidate & Supervisor	i
Acknowledgement	ii
Abstract	iii
List of Figures	iv
List of Tables	vi
List of Abbreviations	vii
Table of Contents	viii
1 Introduction	1
1.1 Transportation Forecasting	2
1.2 Introduction to Mobile Network Big Data	5
1.3 Purpose of the Research	7
1.4 Research Gap	7
1.5 Research Objectives	8
1.6 Scope of the Research	8
1.7 Project Contributions	8
2 Literature Survey	10
2.1 Introduction to Digital Data in Transport Analysis	10
2.2 CDR and Human Mobility	12
2.2.1 Trip Based Analysis Using CDR	12
2.2.2 Use of CDR in Sri Lankan Context	14
2.2.3 Limitations of Using CDR in Travel Predictions	17
2.2.4 Techniques Used in Minimizing the Limitations of CDR	19
2.2.5 Research Gap Identification	20
3 Methodology	21
3.1 Data	21
3.1.1 Call Detail Records	21
3.1.2 Voluntarily Collected Mobile App Data	22
3.1.3 Household Visit Survey (HVS) Data	28

3.2	Preprocessing	30
3.3	Load Sharing Effect on Call Detail Record Data	31
3.3.1	Load Sharing Effect Identification	32
3.4	User Localization	35
3.5	User Activity Pattern Recognition	37
3.5.1	User Profiling Based on Mobility	37
3.5.2	Occupation Identification From CDR Data	38
3.5.3	Feature Engineering	39
3.5.4	Activity Recognition From CDR Data	47
4	Results	51
4.1	User Profiling	51
4.2	Analysis on Load Sharing Effect Identification	51
4.3	User Localization	53
4.4	Activity Recognition	56
5	Conclusion and Future Work	57
5.1	Research Summary	57
5.1.1	Methodological Framework of Using CDR	57
5.1.2	Labeling the Secondary Activities	58
5.1.3	Limitations of the Study	58
5.2	Future Research	59
	References	60