Vehicular Movement Patterns: A Sequential Patterns Data Mining Approach Towards Vehicular Route Prediction

by

Amar Farouk Merah

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the M.Sc. degree in
Electrical and Computer Engineering

School of Electrical Engineering and Computer Science
Faculty of Electrical and Computer Engineering
University of Ottawa

Abstract

Behavioral patterns prediction in the context of Vehicular Ad hoc Networks (VANETs) has been receiving increasing attention due to enabling on-demand, intelligent traffic analysis and response to real-time traffic issues. One of these patterns, sequential patterns, are a type of behavioral patterns that describe the occurence of events in a timely-ordered fashion. In the context of VANETs, these events are defined as an ordered list of road segments traversed by vehicles during their trips from a starting point to their final intended destination, forming a vehicular path. Due to their predictable nature, undertaken vehicular paths can be exploited to extract the paths that are considered frequent. From the extracted frequent paths through data mining, the probability that a vehicular path will take a certain direction is obtained. However, in order to achieve this, samples of vehicular paths need to be initially collected over periods of time in order to be data-mined accordingly. In this thesis, a new set of formal definitions depicting vehicular paths as sequential patterns is described. Also, five novel communication schemes have been designed and implemented under a simulated environment to collect vehicular paths; such schemes are classified under two categories: Road Side Unit-Triggered (RSU-Triggered) and Vehicle-Triggered. After collection, extracted frequent paths are obtained through data mining, and the probability of these frequent paths is measured. In order to evaluate the efficiency and effectiveness of the proposed schemes, extensive experimental analysis has been realized. From the results, two of the Vehicle-Triggered schemes, VTB-FP and VTRD-FP, have improved the vehicular path collection operation in terms of communication cost and latency over others. In terms of reliability, the Vehicle-Triggered schemes achieved a higher success rate than the RSU-Triggered scheme. Finally, frequent vehicular movement patterns have been effectively extracted from the collected vehicular paths according to a user-defined threshold and the confidence of generated movement rules have been measured. From the analysis, it was clear that the user-defined threshold needs to be set accordingly in order to not discard important vehicular movement patterns.

Acknowledgements

First and foremost, I would like to thank Allah Almighty for the endurance, patience, perserverance and strength He has provided me to complete this important chapter of my life.

I would also like to offer my most profound and sincere thanks to Professor Azzedine Boukerche who has provided me with priceless and invaluable support and guidance towards the completion of this thesis. Working under his supervision has been a great experience and a great eye-opener to the great world of research. Without him this thesis would not have been possible.

For their strong belief in me, their endless support, and their continuous encouragement, there are no words that can describe the gratitude I have for my parents. I hope I can continue to make them both proud with all the different endeavors of my life. I also want to thank my brothers, Mohamed and Youcef, and my sister, Ikram, for their moral support and wish them all the best in their academic and professional lives.

I would also like to thank Dr. Samer Samarah for his mentorship, guidance and friendship. I have learned a lot from him throughout the completion of this work, and I hope I can learn more and more.

For all my friends and colleagues in the PARADISE lab, thank you for making this one of the most fruitful, interesting and fun experiences I have ever had. I don't think I would have had a better time in any other research lab, and I hope I can keep on working with you all in the future, whether it be under academic or professional environments.

Lastly, but definitely not least, I would like to thank my best friend, my partner in life and my fiancee, Mouna Achoui, for all the moral support and happiness she has provided me with during this long and hard journey. Thank you for being there for me when I most needed it and for always believing in me. 459 :)

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Recent advances in wireless technology and the cost effectiveness of wireless transceivers have allowed for the emerging of a new class of network architectures that has not been realized before. One of these architectures is the Vehicular Ad-hoc Networks (VANETs). VANETs are a somewhat upgraded and modified version of Mobile Ad-hoc Networks (MANETs) since they also involve moving objects exchanging information with each other. The main idea behind communication in VANETs is to provide vehicles with the ability to exchange information among each other and/or with other central processing units deployed in the network (e.g. on the roadside). The former is known as Vehicle-to-Vehicle (V2V) communication and employs an infrastructure-less approach where vehicles can communicate with each other. The latter is known as Vehicle-to-Infrastructure (V2I) communication and is considered as the infrastructure-based design for VANETs, where vehicles can exchange information with other non-vehicular entities such as traffic lights, road-side units (RSUs), etc. [23] The U.S Federal Communication Commission (FCC) has assigned a 75 MHZ of spectrum in the frequency range from 5.850 to 5.925 GHZ for vehicular communication. This spectrum is divided into seven 10MHZ channels as shown in Figure 1.1. Four channels are for services, one is for controlling, and two are reserved for future advanced applications. [21]

In VANETs, each vehicle is usually equipped with a wireless device consisting of mainly a transceiver and a low-cost Global Positioning System (GPS), collectively known

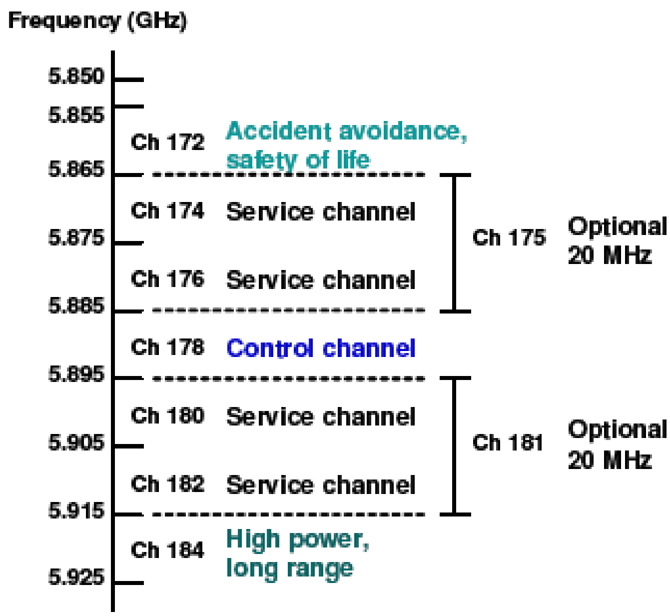

Figure 1.1: The FCC VANET Spectrum [21]

as an On-Board Unit (OBU), which allows it to communicate with other vehicles and infrastructure entities within its transmission range. Thanks to the availability and cost-effectiveness of different communication devices, VANETs have become a future vision that is quickly being realized and has opened the doors to the introduction of a wide range of applications and services to drivers and passengers and are mainly classified under (1) *safety*, (2) *traffic efficiency*, and (3)*infotainment and business* applications and services. [2] [17] Examples for each application classification include:

- ***Safety***: collision avoidance and warning, road condition monitoring, traffic violations, vehicular location tracking, etc.
- ***Traffic Efficiency***: congestion control, traffic monitoring, automated electronic tolling, etc.
- ***Infotainment and Business***: Internet access, music, videos, commercials, driving assistance, parking availability, video conferencing, etc.

However, there are many challenges and issues that have surfaced with the advancement of VANETs and the aforementioned applications. The presence of these challenges and issues is mainly due to the nature of VANETs and has provided researchers with many open areas to discover and explore intensively. There are many characteristics associated to VANETs that are responsible for this and include (1) *dynamic network topology* [1], (2) *decentralized design*, (3) *environmental obstacles and attenuators* (e.g., buildings, trees, other vehicles, etc.). These characteristics, and others, need to be taken into consideration when designing and deploying applications for VANETs, as they can adversely affect the performance of the applications, or be taken advantage of in order to enhance them.[22] [2]

1.1 Motivation

One of the most recent and upcoming applications of VANETs is the prediction and forecasting of vehicular movement behavior [3] [4] [5] [6]. More specifically, the use of behavioral patterns in the prediction of events has been looked at in many different fields and areas such as customer purchasing behavior, financial stocks, medical drug treatment [8], and even robotic movement behavior [7]. Extracting these patterns can be used to better understand the behavior of certain events based on the occurence of previous events happening in a certain manner. One type of behavioral patterns is represented as timely-ordered sets of events and is known as sequential patterns. The main characteristic of this type of patterns lies in the fact that events are sequenced in the order of their occurence, as opposed to others that mainly require events to be present regardless of their order, either solely or with the combination of other events. Based on the historical behavior of the extracted sequential patterns and the current state of the event being monitored, an insight on the future plan of a certain event can be be obtained from the sequence of actions that event has taken to reach where it currently is.

Since the mobility of vehicles differs from MANETs in such a way that it is more predictable than the movement of mobile nodes in a Wireless Sensor Network (WSN), sequential patterns can be exploited to predict vehicular movement behavior [23]. This application has been targeted for many different advantages including, but not limited to, traffic jam avoidance, driver-assistance, and the vehicular pursuit and tracking of criminals. Taking the criminal pursuit scenario as an example, prior knowledge of a fugitive's driving behavior can provide authorities with sufficient information as to where this fugitive may be headed. Using this information, strategic and efficient deployment of road blocks and backup can be achieved, which can significantly lower the time and cost of apprehending such fugitives. Another example, related to traffic efficiency, is the use of previous road congestion information to forecast the level of congestion a certain area or road may suffer from depending on the time of day. Based on this knowledge, drivers could choose alternative less congested routes to reach their destination while avoiding traffic jams and bottlenecks.

In order to extract these sequential patterns and study their behavior, previous historical data needs to be available. For VANETs, there needs to be a set communication schemes that are capable of collecting information from vehicles about the paths they have undertaken during their trips. These communication schemes need to be able to collect trip information in the form of sequentially and timely ordered events in order to make use of sequential patterns data mining.

1.2 Problem Statement and Objectives

In this thesis, the main objective is to design and implement a number of communication schemes that are responsible for collecting vehicular trip information from vehicles on the road and prepare them for sequential data mining purposes. Using the communication schemes, vehicular movement patterns will be obtained and frequent patterns will be extracted to study the effect of sequential patterns data mining on vehicular prediction. Figure 1.2 shows a roadmap of what would need to be done in order to fulfill the requirements of route prediction using sequential patterns.

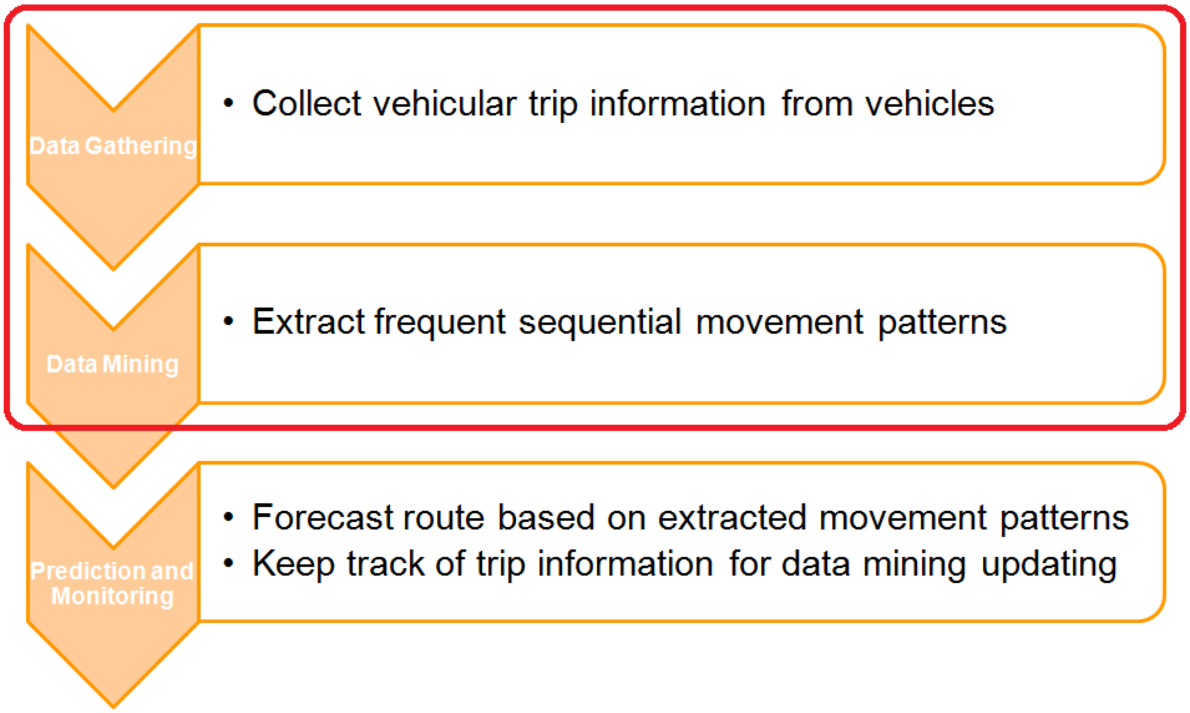

Figure 1.2: Problem Definition Roadmap

This thesis will mainly concentrate on the highlighted parts of Figure 1.2. To better understand each stage, the following definitions are provided:

Data Gathering: In this stage, vehicular movement patterns will be collected from vehicles traveling in a given geographical area in the form of timely-ordered sequences of road segments traversed.

Data Mining: From the collected vehicular movement patterns, sequential patterns data mining will be performed to extract the most frequent movement patterns and measure the confidence of the generated rules to assist in the prediction process.

Prediction and Monitoring: After extracting the most frequent vehicular movement patterns and gaining information about the confidence of the generated movement rules, prediction of vehicular movement behavior can be achieved. Using the collected vehicular paths in the ***Data Gathering*** stage alone is not enough, and there is a need to continuously gather more paths in order to extract more frequent movement patterns and generate more movement rules, which will provide a stronger prediction accuracy.

In more specific terms, this thesis will cover the following objectives:

- Perform a comprehensive study of current vehicular trip information gathering mechanisms, data mining techniques and prediction models and methods used in VANETs. Each related work will be explained and the application it serves will be provided. A summary of each topic will be provided illustrating a comparison that will highlight their strengths and weaknesses.

- Redefine sequential patterns generic definitions to fit the context of VANETs. Each definition will be constructed with respect to the data mining aspects of sequential patterns.

- Design and implement novel communication schemes that will be used to gather vehicular trip information from vehicles traveling in a given geographical map. These communication schemes need to be able to collect accurate vehicular paths in such a way that does not burden the network with communication overhead and latency.

1.3 Thesis Contributions

In terms of contributions, the following will be provided in this thesis:

- A set of formal definitions representing vehicular movement behavior in the form of sequential patterns. Important data mining definitions from [27] will be tailored to fit the context of their operation in VANETs.

- Five novel communication schemes that make use of the infrastructure on the road to collect vehicular paths. The operation of these communication schemes will be explained in details, illustrated in an algorithmic form to provide a clearer and more systematic representation of the schemes, and a performance evaluation in terms of network efficiency will be analyzed and discussed from results achieved through an extensive set of experiments.

- An analysis of the data mining results performed on the extracted frequent vehicular movement patterns in terms of minimum support and confidence of the generated rules.

1.4 Thesis Organization

This thesis is organized as follows. Chapter 2 discusses the related works in the area of VANETs and data mining techniques that have been used to predict movement patterns. It will discuss the different studies and techniques that have been used to gather information about vehicles paths that have been undertaken during a certain period of time, different techniques used to extract useful driver behavior information which can be used for prediction. Also, a survey of different prediction techniques to forecast vehicular movement will be provided.

Chapter 3 provides some formal definitions which describe vehicular movement patterns in terms of sequential patterns. Some detailed examples will also be provided for these definitions. In Chapter 4, a full detailed description of the path information gathering mechanisms will be provided, including an algorithm to better understand each mechanism.

Chapter 5 will present a performance evaluation and analysis to display the strengths and weaknesses of these mechanisms, the tools and parameters used for the simulations carried out and a discussion of the obtained results, as well as a presentation on how frequent sequential patterns generated from vehicular movement patterns can be used to predict the movement of vehicles in a VANET. Chapter 6 will conclude the thesis and highlight what has been achieved, as well as propose some further improvements which could be studied and implemented as future works in the same area.

Chapter 2

Literature Review

In this chapter, a survey of different works and studies related to the topic of the thesis will be discussed. More specifically, this literature review will encompass the aspects of vehicular trip information gathering techniques, both in earlier and current studies, several data mining techniques and their applications, and finally mobility prediction techniques, mainly in VANETs, used to forecast the movement behavior of mobile entities.

2.1 VANETs and Vehicular Path Gathering Techniques - Earlier Studies

Departments of transportation and many employers have, since a long time, been interested in the trajectories vehicles undertake during their journeys, as well as other factors. They may have needed this information for many different reasons which include, but are not limited to, efficient analysis of traffic conditions throughout different times of the day, congestion and traffic control in urban areas and proper freight modeling and planning for on-time delivery of commodities. Some of the earlier methods of collecting information about vehicular trips were used by truck delivery companies in order to better track the movement of goods and find more efficient ways to deliver these goods.

[9] In this section, three different early trip information gathering methodologies will be discussed; mail-out mail-back surveys, telephone surveys and roadside interviews.

2.1.1 Mail-out Mail-back Surveys

With the heavy and long schedules truck drivers undergo during their shifts, it would be hard to find the time to acquire information about the trips they have undertaken. A survey was created in order to collect this information from the drivers and would be sent to them to fill at a time most convenient to them. In 1986, the Chicago Area Transportation Study for Commercial Vehicle Survey sent out surveys to truck drivers by mail in order to better model and plan policies in the northeastern Illinois area. The data collected from the survey included information such as trip frequency, distance, purpose and land type use over a 24-hour period and included three different types of vehicles: light, medium and heavy. They received a complete response rate of approximately 25% of the randomly chosen sample of 17,834 drivers. [28]

Cambridge Systematics, Inc. performed a mail-out mail-back survey for commercial vehicles and truck traffic within Phoenix, Arizona metropolitan area in 1991 for the Arizona Department of Transportation. This survey covered trip information specifics such as origin and destination addresses, vehicle type, estimated gross weight, odometers readings and vehicle usage for a 24-hour period. They were able to achieve a response rate of 30%. [29]

2.1.2 Telephone Surveys

In 1994, the Metropolitan Planning Organization for the city of El Paso, Texas and the Texas Department of Transportation performed a telephone survey for the El Paso metropolitan area and were concerned mostly with travel demands and air quality modeling, and included information such as origins and destinations, trip purposes and route information for each trip segment. The response for their survey achieved a rate of 43%.

[30]

Also in 1994, Wilbur Smith Associates performed a combination of telephone and mail-out and mail-back surveys for the Houston-Galveston Area Council in order to capture truck travel characteristics from vehicle owners/operators such as truck type, odometer readings and origin-destination information. Their random sample, obtained from the Department of Motor Vehicle, included other commercial vehicles as well and their estimated response rate ranged between 35%-40%. [31]

2.1.3 Roadside/Tollbooth Interviews

In the period between 1974 and 1991, the Port Authority of New York and New Jersey deployed roadside surveys at a number of tollbooths across New York and New Jersey which were designed to monitor and track the movement of commodities through highway corridors by truck drivers. A number of attributes were included in the survey which included truckload type, origin and destination, commodities on board, travel stops, etc. [32]

Two other places where roadside interviews were used are with Ontario's Ministry of Transportation, Canada from 1978 to 1988 and Alameda County, San Francisco, California in 1991. Ontario's Ministry of Transportation collected information from truck drivers on a 5-year interval rate and this information was used for corridor analysis and better planning of inter-city movement policies, and collected information such as origin and destination, vehicle type, commodity and vehicle weight. [34] For the Alameda County, Barton-Aschman Associates conducted four different surveys to better manage the movement of commodities by truck drivers of three classes of trucks. For the roadside interviews, they were dispersed over a number of weigh stations and toll bridge crossings, and information was collected at these points. [33]

2.1.4 Comparison of Early Vehicular Trip Gathering Techniques

In summary of these three trip information gathering techniques, it is apparent that the response rate for the mail-out mail-back and telephone survey techniques is not as high as might be needed, reaching a high of 43% of complete responses. Also, information received through these techniques may not be as accurate as expected due to the fact that it depends on the drivers' input, which is prone to human error and bias. The roadside/tollbooth interview technique might be able to achieve a higher response rate and accuracy, but at the price of the labor cost associated with the deployment of interviewers and disruption of traffic to perform the interviews. [9] Table 2.1 displays a summary comparison of the aforementioned early studies on vehicular trip information gathering techniques.

Survey Technique	**Cost**	**Response Rate**	**Accuracy**	**Road Disruption**
Mail-out Mail-back	Low	Low	Low	No
Telephone	Low	High	Low	No
Roadside/Tollbooth Interviews	High	High	Moderate	Yes

Table 2.1: Summary Comparison of Early Vehicular Trip Gathering Techniques

2.2 VANETs and Vehicular Path Gathering Techniques - Current Studies

Many different technologies have emerged since the previously mentioned earlier studies that do not require human interaction or input and can collect data automatically without the need to disrupt traffic. Some applications may need accurate real-time information, whereas others can accommodate for error margins in the accuracy. [10] In this section, some of the recent techniques used to localize vehicles in a certain geographical area will be reviewed. They may not all be trip information gathering techniques, per se, however can be used to gain knowledge of a vehicle's whereabouts and the trips it may have undertaken during a certain period of time.

2.2.1 GPS, DGPS and Map Matching

Global Positioning Systems (GPS) have been used to provide information on the location of objects that are equipped with a GPS receiver. With the help of 24 strategically placed satellites in orbit around the Earth, GPS-compatible objects (stationary or moving) can be found, provided there is a direct line of sight with the satellites. By receiving information from different satellites, vehicles can estimate their position according to this received information with techniques such as Time of Arrival (ToA) and trilateration. However, this information is not accurate due to the known localization error in GPS receivers, which can reach up to 30m. [10] This can be an issue for critical and emergency types of VANET applications which need accurate location. Another issue with GPS is the line of sight constraint, and this occurs when obstacles such as buildings, tunnels, or trees for example obstruct or maybe interfere with the direct line of sight GPS receivers require in order to receive GPS information. [11] This could also render the accuracy inapplicable for VANET applications that require accurate location information.

In order to handle the localization error problems associated with GPS, an updated technique has been introduced which makes use of a stationary GPS receiver with a known location and similar localization errors as other receivers. This GPS-based localization technique, known as Differential GPS (DGPS), provides moving vehicles with the differential error information between the actual position of a fixed GPS receiver and that which is received from a satellite for that same receiver. Once this error is calculated by the fixed GPS receiver, it broadcasts it to all other vehicles in order to synchronize their receivers accordingly. [10]

Map matching has been considered as a tool used to enhance the localization of a moving vehicle and not as an actual localization technique. By measuring the position of a moving vehicle using a localization device (e.g. GPS), map matching can find this location on a preloaded map and pinpoint the location of the vehicle on those maps.

Collecting different location points over a period of time can give the user an approximate idea of the trip that vehicle has been through. [10]

2.2.2 Cellular Localization and Tracking

The authors in [12] and [14] have proposed the use of wireless ATM cellular networks to track the movement of mobile users through a geographical area. Using a cellular network infrastructure and the signal received from mobile devices, a sequence of ATM cells traversed can be obtained. Figure 2.1 shows an illustration of two users moving through a cellular network.

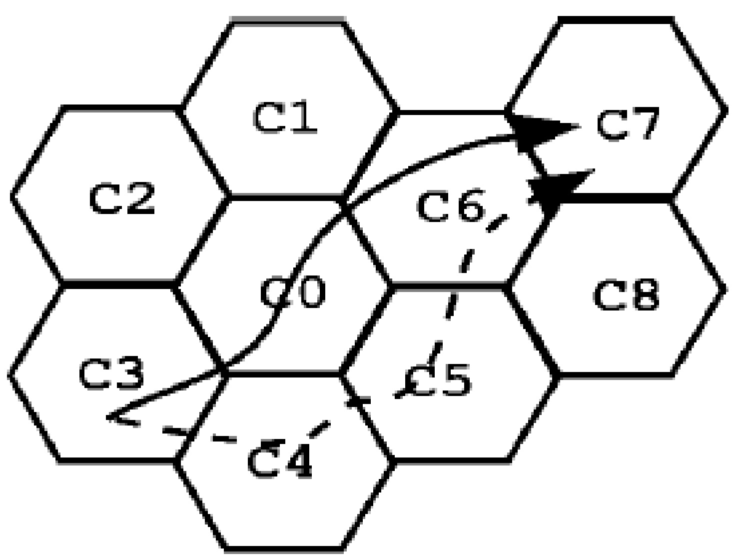

Figure 2.1: Mobility Tracking using Cellular Networks [12]

For the user represented by the solid line, the sequence of ATM cells traversed would be: [C3, C0, C6, C7] and [C3, C4, C5, C6, C7] for the user depicted by the dotted line. This localization is possible due to the need for mobile devices to exchange information with cellular towers during mobile handoff processes for example. In VANETs, the authors in [13] took advantage of the cellular handoff activity of drivers' mobile devices, in addition to video surveillance, to estimate the congestion level of traffic in a certain geographical area. For an application like congestion detection, cellular activity is a reasonable solution to be employed which can give an approximate idea of how much traffic is in a certain area. However, specific road segments will be hard to detect since cellular localization is based on the location of the cellular tower itself and all road segments within its transmission range will be considered as one area traversed.

2.2.3 Client-Server Location Tracking Update Policies

In order to keep track of moving vehicles, client-server update policies for VANETs have been discussed in [15] which fall under three categories: (1) *point-based*, (2) *vector-based*, and (3) *segment-based*. These update policies consider vehicles as clients and a central database as the server, and assumes a wireless communication network is in place to provide location updates from the vehicles to the central database. Figure 2.2 provides a representation of each update policy.

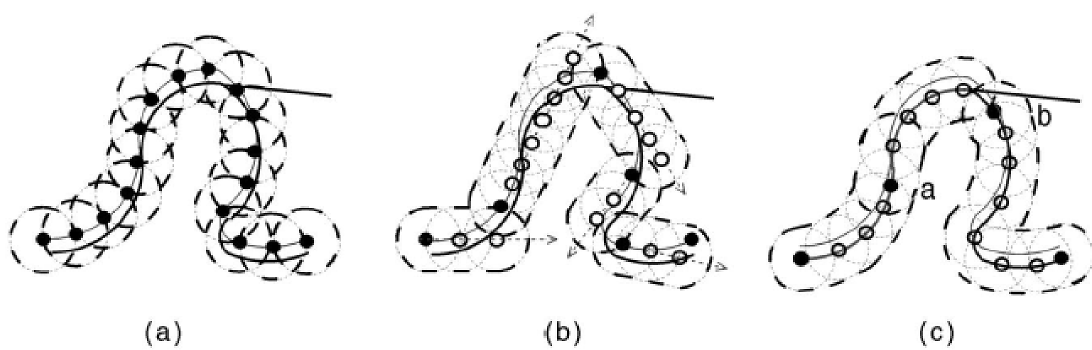

Figure 2.2: Tracking Update Policies: (a) Point-based, (b) Vector-based, and (c) Segment-based [15]

As can be seen in Figure 2.2(a), the vehicle updates its location to the server every time it crosses a certain distance threshold. Each solid point in the figure represents an update event and the aura around that point is the distance threshold for that point. In Figure 2.2(b) however, the location update is based on a linear function of time, where both the start point and velocity of the moving vehicle are taken into consideration. In the same figure, the solid points represent moments at which the vehicle sent an update to the server and the hollow points represent predicted positions by the server, since these tracking update policies were designed for that purpose. Finally, Figure 2.2(c) shows the operation of the segment-based approach, where vehicles will update the server with their location based on the road segment they are about to traverse. Given that the map knowledge is provided to the vehicle and GPS coordinates information as well, map matching can be used to detect the presence of a new road segment and hence update the server accordingly. The solid points in Figure 2.2(c) signify the entrance of the vehicle into a new road segment, whereas the hollow points, just like in Figure 2.2(b) signify predicted points by the server.

2.2.4 Comparison of Current Vehicular Trip Gathering Techniques

In this section, some techniques used to provide location information to vehicles and other entities such as central databases have been explained. GPS, DGPS, and map matching have been able to provide location information to vehicles to a certain accuracy and with some delay due to transmission latency from vehicles to satellites. Cellular localization can be used to gain information about the approximate number of vehicles in a certain geographic area but is constrained by the size of each cellular cell and the availability of mobile devices. This limits its capability to provide accurate road traffic information and precise vehicle location in terms of road segments traversed. The client-server update policies described provide an interesting approach towards a more specific and precise location tracking. The point-based and vector-based update policies may be a little too costly in terms of frequency of updates, whereas the segment-based approach provides a lesser number of updates since it is only required to provide the server with its new location once a new road segment is detected. Table 2.2 provides a short summary of the cellular and client-server location tracking update policies. The accuracy and issue column represent the reason why these policies might not be suitable for the purpose of sequential data gathering. If the accuracy is low, for example, this means that the location information is not enough to concisely specify which road segments have been traversed by a vehicle.

Location Update Policy	Update Frequency	Accuracy	Data Reported	Issue
Cellular -Based	Low	Low	Cellular area activity	Many road segments in one cell
Point -Based	High	High	Points on a road segment based on distance	Many unnecessary updates
Vector -Based	Moderate	Moderate	Location based on velocity and start point	Some unnecessary updates
Segment -Based	Moderate	High	Location based on segment detection	Some unnecessary updates

Table 2.2: Summary Comparison of Current Vehicular Trip Gathering Techniques

2.3 Data Mining Techniques in VANETs

This section aims at reviewing the different data mining techniques used in VANETs and the applications they have served when implemented. Most of these data mining techniques have been used previously in other applications and industries, but mainly focus on predicting the future behavior of a certain entity based on already available data. Since the topic of this thesis is related to the field of VANETs, this discussion will only look at those data mining techniques that have been implemented for the sake of applications in VANETs.

2.3.1 Association Rules

Association rules have been used to extract information about the occurence of certain events based on a minimum support value, which basically provides a criteria for how sensitive the data mining process will be. [35]. These association rules can be used to discover common attributes or relationships between events or objects, as long as they appear within the same context of a given data collection [36]. In [16], the authors have made use of association rules to detect the presence of faulty and malicious vehicles on the road based on the occurence of abnormalities in their behavior. In this work, vehicles communicate with each other in a peer-to-peer manner and share safety information and warning messages. These messages are collected and association rules data mining is performed to in order to prevent malicious or malfunctioning vehicles to participate with false information that might deteriorate the accuracy of the data collected. Based on the collected temporal correlation rules between vehicles behavior and their computed confidence, the mechanism performs an analysis and detects which vehicles behave with anomalies compared to previously collected correlation rules between vehicles.

Bae and Olariu in [20] have also taken advantage of association rules mining in an application for VANETs. In their work, association rules mining was used to extract control rules for the linguistic information system in their proposed context-aware driving assistance system that helps prevent the occurence of traffic accidents. Based on current vehicle situation information and the extracted control rules, a pattern similarity mechanism is used to find a correlation between those current events and those in the information system that have resulted in accidents or fatalities. Using this mechanism, the system can provide drivers with the most effective and safe actions to be taken in order to avoid an accident or reduce the effect of such an accident occuring.

2.3.2 Classification

In this type of data mining technique, a model is constructed which describes a set of predetermined classes from a set of tuples known as the training set. This data mining technique aims to define the grouping of unknown objects based on some of the attributes of this unknown object. This technique is supervised, which means that it requires an input from the user, in this case the training set, to build the model and train it accordingly. [37] In the work described in [18], the authors aimed at providing a security engineering approach for VANETs towards the assessment of the security needed for certain applications. In order to do so, the authors proposed the use of classification data mining in order to analyze the large set of VANETs applications, classify them according to their security requirements (e.g., severe authentication level required, replay attack susceptibility, etc.) and provide security solution for each class of application. Using the model constructed, new application can be assessed based on their security requirements and the appropriate security measures can be applied.

2.3.3 Clustering

Similar to classification data mining, clustering aims at grouping objects of similar attributes into classes. This technique however is unsupervised, and therefore does not require manual input of a set of sample classes that can aid in the model construction. [37] The work in [19] aims at aggregating similar or closely similar abnormal recordings of speed from vehicles in a cluster in order to provide discovery and dissemination of congestion information. In this approach, the vehicles have taken the responsibility of performing the analysis on the collected data as opposed to the traditional way of burdening a single centralized database to do so. Their system was able to construct clusters in a disconnected V2V environment using two methods: (1) *K-Means* and (2) *agglomerative clustering*. The former method, which constructed the clusters based on the weight of the source data, was found not to perform well when specifying apriori cluster numbers, whereas the latter was able to find clusters without the need of apriori cluster numbers.

2.3.4 Summary

To the best of my knowledge, these were the main works in the literature that made use of different data mining techniques to ameliorate the performance of certain applications in VANETs. There are many more data mining techniques that have been used for different applications and industries, but the aforementioned studies were specific to VANETs. As can be seen, no work has been done to extract vehicular trip information in a sequential patterns data mining fashion.

2.4 Mobility and Route Prediction

Many works have looked at the different techniques to predict the movement behavior of mobile nodes and vehicles in order to forecast their destination, reduce congestion levels in certain geographic areas, etc. This section will provide some of the most recent works researched and proposed to discover vehicular future movement and destination. The work described here is mostly related to VANETs, with the exception of the last technique which is based on the prediction of mobile nodes in a Wireless Sensor Networks (WSNs).

2.4.1 Simple Markov Model

In the work presented in [3], the authors proposed the use of a Simple Markov Model (SMM) that is built based on the observation of previously traversed road segments to predict near term future routes. Using two variables to define each traversed road segment, time of traversal and road segment traversed, the SMM is able to provide a probabilistic forecast of the next road segment. This prediction accuracy though is dependent and sensitive to the time of day in which the current traveling vehicle is trying to predict the next road segment as well as the number of past segments observed. In their experiments, to evaluate the accuracy of their technique, the authors compared their SMM to two types of random guessing algorithms; one that guesses the next turn to take at an intersection randomly with knowledge of the driver's intended direction and one that also guesses the next road segment based on no direction of travel being known. In their results, for predicting one segment, it showed that the SMM was able to attain a prediction accuracy of 90% as compared to direction known random guessing which achieved 50% and direction unknown random guessing with 25% prediction accuracy. The accuracy of prediction dropped even lower for each technique as the number of segments predicted in the future increased.

2.4.2 Hidden Markov Model

Simmons et al., in [6], presented a predictive approach for learning the turning intent of drivers as they travel in a geographic area. Using a low-cost GPS and map matching through a pre-loaded map, a Hidden Markov Model (HMM) was constructed which uses past vehicular behavior to predict the next road segment to be taken in a real-time manner. This model is built to track processes (i.e., driver's actions) with hidden states (i.e., destination and route), and generally needs vast amounts of data to during its learning phase. The advantage here is that the already known map can be used to define the structure of the model, and the number of states is limited to the number of decision a driver must make at an intersection. In addition, the HMM was able to include other factors such as vehicular speed, time-of-day and day-of-week to its prediction model, making the prediction more realistic and accurate. However, the usage of continuous state parameters can prove to be an issue for learning HMMs, and therefore these factors have been described as discrete values with each value representing a range of values. For instance, the time-of-day parameter could be defined as morning rush hour for traveling times between 8AM and 10AM , 10AM-12PM as late morning, 12PM-4PM as early afternoon, etc. However, when new factors are added, the HMM needs to go through a re-learning phase to include all of these new factors. The authors claimed an accuracy of 98% in their work for most of the cases tested, but this was achieved with vehicles not having too many choices to make, as 95% of the cases included vehicles having to choose from only one road segment option, as opposed to more than one as is the realistic case in an urban environment [25].

2.4.3 Variable-order Markov Model and Probabilistic Suffix Trees

In [24], the authors used real taxi GPS mobility trace data to define the mobility patterns of their trips. Using a Variable-order Markov Model (VMM), the data was mined and significant mobility patterns were extracted and variable road conditions were included in the model training to provide a more realistic definition of these patterns. The patterns were selected based on a certain probability of occurence threshold, in order to include only those patterns that occur within a confidence threshold, and a Probabilistic Suffix Tree (PST) was constructed from these patterns to predict the next road segment to be traversed by a vehicle. In their experiments, the authors compared the prediction accuracy for the same technique with the inclusion of variable traffic conditions and without it. In the former, the accuracy seemed to be higher by an average of 7% than the prediction performed without consideration of traffic condition.

2.4.4 Closest Match Prediction

The authors in [25] also made use of previous GPS traces and the fact that many drivers follow frequently repeated routes to their destinations in order to make a long-term prediction of a vehicle's trip trajectory. In this work, Froelich and Krumm proposed the use of full route matching and closest route matching to compare the routes of currently traveling vehicles with historically collected and repeated routes. The techniques provided an ordered list of the routes most similar to the current trip from the historical routes based on some factors including start position, end position and direction of traveling. For the full route matching, the prediction was not useful or accurate for cases where different start positions, end positions and directions were used. For instance, if a vehicle followed a certain route during its trip but had a final destination that was one road segment ahead than the final destination of a previously collected, this historical route will be disregarded. In the closest route matching, a trip similarity measure is used to select the best matching routes to the one currently being traveled. The prediction accuracy was measured as the number of correctly predicted routes with respect to the percentage of trip completed as well as the number of miles driven. The prediction recorded was pretty low (i.e., 40% when half of the trip has been completed), but the correct route would occur 97% of the time in the top ten matches provided by the closest route matching technique.

2.4.5 Prediction-Based Tracking Technique Using Sequential Patterns

To the best of my knowledge, the use of sequential patterns data mining has not been used to extract vehicular movement patterns in VANETs. In WSNs however, Samarah et al., proposed a sequential pattern based prediction technique (PTSP) in [26] to track and monitor moving objects using sensors in order to predict their future position. The mobility patterns of the mobile objects were collected in a sequential manner depending on their time of passing by a sensor. The frequent patterns were extracted based on a user-define minimum support, and were represented as 3-element patterns called tri-sensor patterns. Each tri-sensor pattern defined three sensor nodes passed, and during the prediction, each future location of a moving mobile object was detected by observing the previous sensor visited by that object and the currently detected sensor. Figure 2.3 illustrates how a full sequential pattern was broken down into tri-sensor patterns for the sake of real-time prediction. In this work, the amount of energy lost was reduced and the missing rate of objects being tracked was kept to an appropriate and acceptable level. In order to achieve these two results, two mechanisms were also introduced; a sensor node activation mechanism and a missing object recovery mechanism. The former was designed to put some sensor nodes to sleep based on the knowledge that those sensor nodes will not have objects moving in their vicinity, therefore saving network energy. The missing object recovery mechanism was designed to detect and find any missing objects that may have gone astray from the predicted route, and this mechanism can be run in four modes; (1) *source*, (2) *destination*, (3) *all neighbors recovery*, and (4) *all network recovery*.

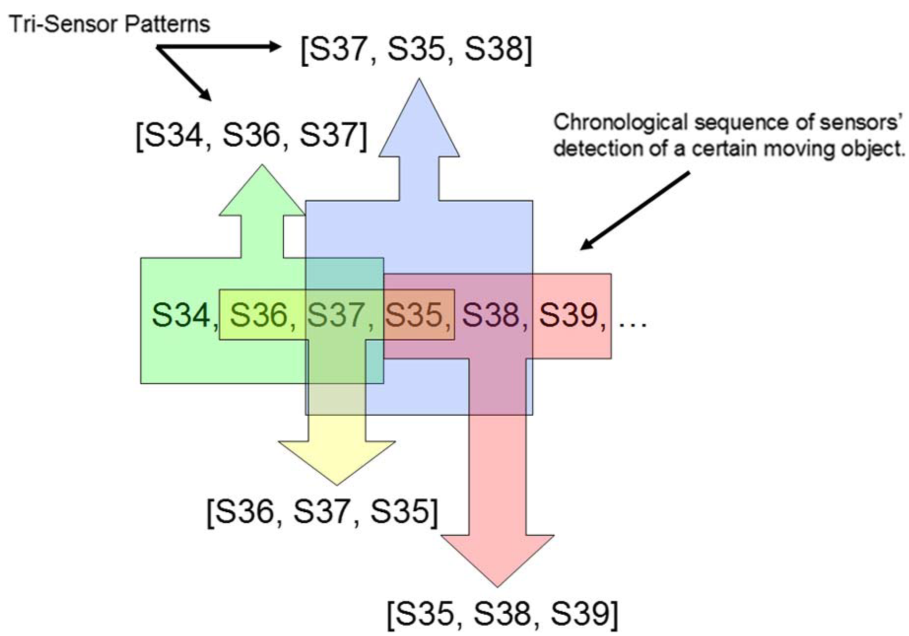

Figure 2.3: Trisensor Pattern Generation [26]

2.4.6 Comparison of Prediction Techniques

In summary, these are some of the most recent and current prediction techniques that have been used to predict mobility of moving objects. For VANETs, and to the best of my knowledge, no work has been done to predict vehicular mobility with the use extracted sequential patterns. Table 2.3 shows a summary of the aforementioned prediction techniques.

Prediction Technique	Accuracy	Predictive Vision	Based on	Accuracy Constraint
SMM	High	Next road	Markov chain	No. of observed segments
HMM	Moderate	Next road	Markov chain	No. of turn decisions
VMM + PST	Moderate	Next road	Markov chain and Probabilistic Trees	Road conditions
Closest Match	Low	Long term	Moderate	Current trip and time
PTSP	Not reported	Next location	Sequential Patterns	Trisensor Patterns

Table 2.3: Summary Comparison of Mobility and Route Prediction Techniques

2.5 Summary

In this chapter, three important topics have been discussed and reviewed; vehicular trip data gathering techniques, data mining techniques and route and mobility prediction techniques, in order to broaden the understanding of what is currently available in the literature. Almost all of the work that has been discussed in this chapter is in direct relation with VANETs, except for the last prediction technique which has been implemented for the sake of WSNs. From the literature review provided above, and to the best of my knowledge, there has not been any work that allows vehicular trip information collection in the form or manner that will be presented in this thesis (i.e., in the form of sequential patterns).

Chapter 3

Vehicular Movement Patterns

As mentioned in Section I, sequential patterns have been used in many different applications. In order to apply the concept of sequential patterns to this work, a set of formal definitions need to be defined with respect to the movement behavior that vehicles in VANET follow. These definitions are inspired by the work presented in [27] and are tailored to represent the vehicular movement patterns in question.

3.1 Movement Patterns Formal Definitions

Definition 1 *Let* $S = \{S_1, S_2, \ldots , S_{|i|}\}$ *represent the set of road segments in a given geographical area or map. A road segment* $\boldsymbol{S_i}$ *is represented by a unidirectional edge between two consecutive junctions. The junctions are also considered as road segments in this work, and can represent not only an intersection on the road, but also the end of the road, exits, etc.*

In order to be better understand Definition 1 provided above, Figure 3.1 illustrates an extract from a geographical map which describes the representation of a road segment and junctions. As can be seen from the figure, a lane going in one direction is considered a road segment, while the lane in the other direction is a different road segment. Also, each junction can connect 2 or more road segments together.

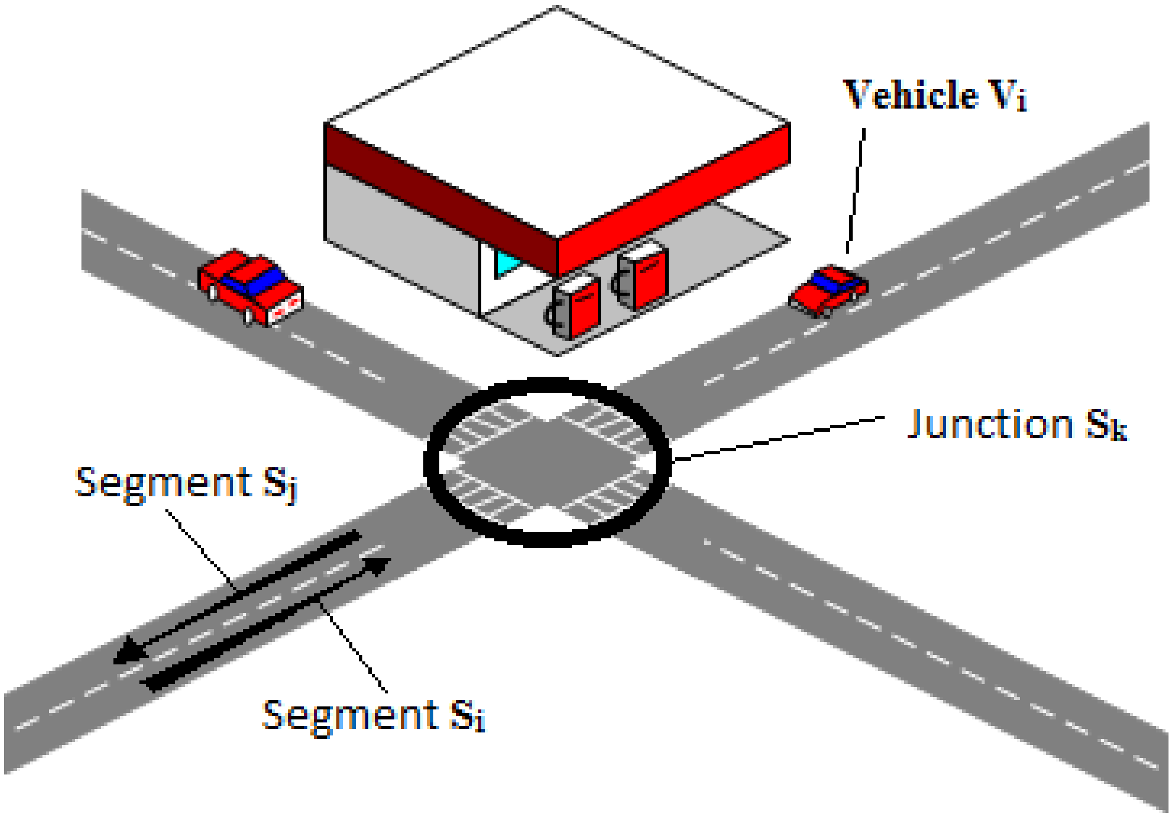

Figure 3.1: Road Segments and Junctions

Definition 2 *Let $V = \{V_1,\ V_2,\ \ldots\ ,\ V_n\}$ be the set of vehicles that are traveling in the given geographical area during a certain time period.*

Definition 3 *Let $M = [S_1,\ S_2,\ \ldots\ ,\ S_n]$ represent a vehicular movement pattern that depicts the road segments traversed by a particular vehicle V_i during its trip in a specific geographic area within the given map.*

The Movement Database, shown in Table 3.1, represents the set of vehicles movement patterns during their trips in that monitored area. As shown below, each vehicle V_i will have an entry in this database describing its movement patterns.

Definition 4 *A movement pattern $M = [S_1,\ S_2,\ \ldots\ ,\ S_n]$ is said to be a sub-pattern of M' if its elements are contained in the exact order as the patterns in M'.*

V_{ID}	**Vehicular Movement Patterns**
V_1	[S_1, S_3, S_5, S_9]
V_2	[S_3, S_8, S_{12}, S_{13}]
...	...
...	...
V_n	[S_1, S_2, ... , S_n]

Table 3.1: Movement Database

This is an important point to take into consideration when dealing with sequential patterns, since they represent patterns of events occuring in a certain order. Sub-patterns must not only contain **all** the elements in the respective movement pattern, but the elements needs to be contained in the **same exact order**. For example, the movement pattern [S_3, S_4, S_6] is said to be a sub-pattern of the patterns [S_1, S_3, S_4, S_6] and [S_0, S_1, S_3, S_4, S_6, S_8], but not a sub-pattern of [S_1, S_3, S_4, S_5, S_6]. Even though the elements are available in the last pattern, they do not follow the same order and therefore cannot be considered as a sub-pattern of it.

Definition 5 *The support of a movement pattern M, represented as* ***Support(M)****, is defined by the number of movement patterns in the Movement Database in which M can be identified as a sub-pattern.*

In sequential patterns, this is an important definition to understand in order to extract frequent sequential patterns from a sequential patterns database (i.e., movement patterns database in this case). The support basically provides the frequency of patterns in the movement patterns database. In order to extract these frequential patterns, a user-defined minimum support threshold is compared with the support of a specific movement pattern ***Support(M)***. If ***Support(M)*** is greater than the given minimum support, then that movement pattern ***M*** is considered a frequent movement pattern. [51]

Definition 6 *If* $M = [S_1, S_2, ..S_n]$ *is a movement pattern:*
A movement rule ***R*** *is defined by the implication that* $M_1 \Longrightarrow M_2$*, where* M_1 *and* M_2 *are*

sub-patterns of M and there are no common elements (road segments) between M_1 and M_2.
*The support of the rule is **Support (M)**.*
*The confidence of the rule **R**, **Conf (R)**, is defined by*

$$Conf(R) = \frac{Support(M)}{Support(M_1)} \tag{3.1}$$

The confidence of a rule here depicts the frequency of a movement pattern occuring with a given sub-pattern. Depending on the minimum support threshold defined, the confidence of a rule can provide an insight on the probability that a specific sub-pattern can occur in a movement pattern given that another sub-pattern occurs in the same movement pattern. For example, in the scenario where a vehicle reaches a junction, based on the movement patterns collected in the database, a confidence can be obtained for each different turn decision (e.g., right turn, left turn, etc.) depending on the previously traversed road segments by that same vehicle. In a more examplary manner, if a vehicle has gone through road segments [S_1, S_3, S_4] before reaching road segment S_6, the confidence of the generated movement rules for each turn decision can be analyzed in order to see how frequently the next road segment was chosen by previous vehicles in the movement patterns database with the same given sub-pattern [S_1, S_3, S_4].

3.2 Summary

In this chapter, we have defined some important terms and definitions to represent vehicular movement in the context of sequential patterns. The movement of vehicles were defined as a sequence of road segments traversed during their trips. The support of a movement patterns was defined also, another important factor when extracting frequent sequential patterns, as well as the confidence of the generated movement rules.

Generating movement patterns requires accumulating the vehicles behavior (i.e. done using the Movement Databases), which will be the focus of the next chapter. Rules will be generated from the patterns that are already recorded in the database. For instance, let us consider the pattern M = [S_1, S_2, S_3, S_4]. The set of possible rules will be

$[S_1] \Longrightarrow [S_2,\ S_3,\ S_4]$

$[S_1,\ S_2] \Longrightarrow [S_3,\ S_4]$

$[S_1,\ S_2,\ S_3] \Longrightarrow [S_4]$

Chapter 4

Vehicular Movement Patterns Data Preparation Schemes

In order to extract frequent movement patterns, relying on historical mobility information collected over a period of time is required [38]. After collecting all the full vehicular paths and determining a minimum support threshold, the most frequent traveled vehicular paths can be extracted and used as vehicular movement patterns [39] [40]. In this chapter, five novel communication schemes responsible for collecting vehicular path information are presented. These schemes have been classified under two categories; RSU-Triggered and Vehicle-Triggered. The main difference between the two is the operation behind the communication that happens between the RSU and vehicles. The vehicle-triggered schemes, namely the Newest Edge and Full Path schemes, are also classified into two modes of operation; Broadcast Mode and RSU Discovery Mode. Figure 4.1 shows an a diagram which illustrates the classification of the schemes proposed in this chapter.

In the implementation of each scheme proposed, RSUs are distributed all over a geographical area in such a way that all road segments are under their combined coverage. Each vehicle will be able to detect a new road segment using a map matching technique, and each RSU will be responsible for keeping track of the partial paths sent by each vehicle in its vicinity. In the next sections to come, each communication schemes will be described in details and an algorithm for each vehicle and RSU will be provided to

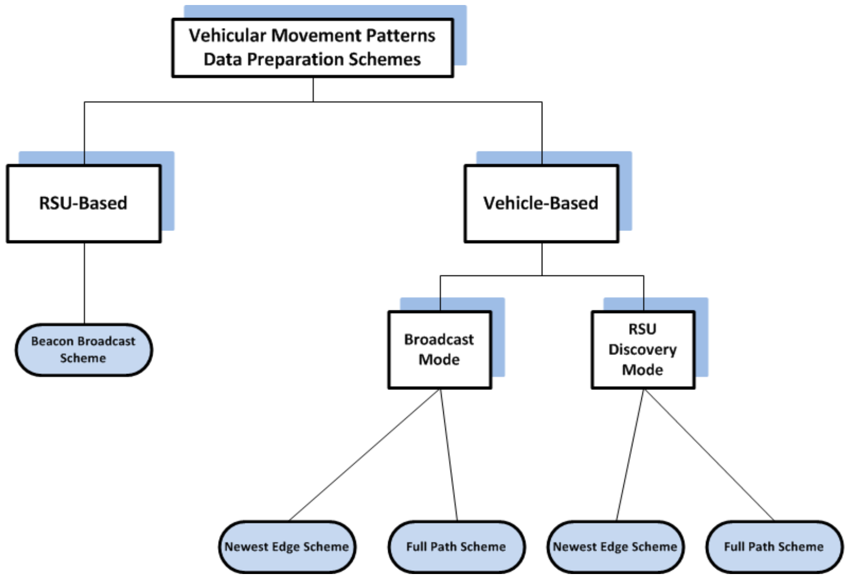

Figure 4.1: Vehicular Movement Patterns Data Preparation Schemes Classification

explain their operation in a clearer and more systematic manner.

4.1 RSU-Triggered Scheme (RT)

In this scheme, vehicles and RSUs both play important roles in the data gathering process. As vehicles move through the VANET, they record every road segment S_i they traverse in the order of arrival to those segments. RSUs are assumed to be located at fixed locations and each RSU covers a certain number of road segments according to their radio transmission coverage. At every X seconds, each RSU will broadcast a beacon to all vehicles in its vicinity, which will query each vehicle to send their collected set of segments, also known as partial paths PP_i. Upon receipt, the vehicles will create a packet containing the partial path collected as well as other attributes shown in Figure 4.2. Each vehicle has a unique identifier V_i and will only send the road segments it has collected between two consecutive beacons. In order to identify which RSU to reply to with the partial path, the *RSUBeacon* message received by the vehicle will contain the orginating RSU's address *RSUAddr*. Algorithm 1 describes the operation of each vehicle in this scheme.

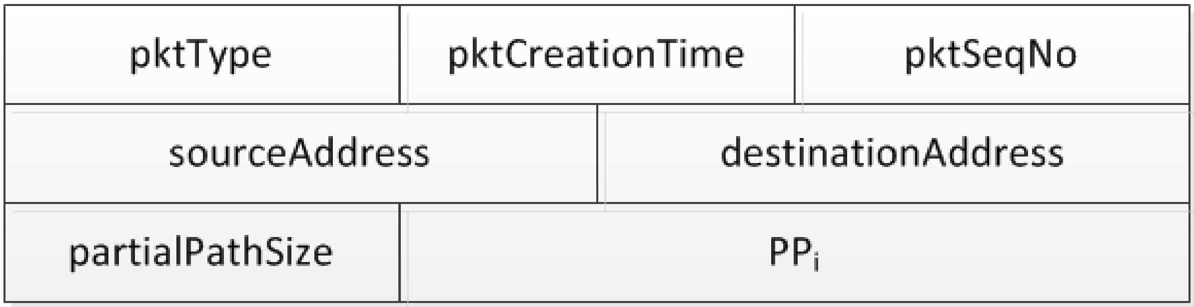

Figure 4.2: Packet design for RSU-Triggered Scheme

Algorithm 1 Vehicle($\mathbf{V}_i$, $\mathbf{PP}_i$[ ])

while true **do**
  get road segment $\mathbf{S}_i$;
  if new $\mathbf{S}_i$ **then**
    append new $\mathbf{S}_i$ to partial path $\mathbf{PP}_i$[ ];
  end if
  if RSUBeacon received **then**
    send ($\mathbf{V}_i$ and $\mathbf{PP}_i$[ ]) to **RSUAddr** from **RSUBeacon** ;
  end if
end while

At the RSU side, received collected partial paths will be added to the RSUs' databases according to the nature of the received packet. If the received PP_i contains a vehicle ID (i.e. sourceAddress in Figure 4.2) that does not exist in the RSU database, the RSU will create a new entry for this vehicle ID and include the associated PP_i to the entry. However if the vehicle ID in the packet already exists, the partial path is appended to the existing vehicle path. At the end of the collection, or when triggered, the RSUs can provide each other with the partial paths collected and use the *pktCreationTime* attribute associated with each PP_i packet received to construct a timely-ordered sequence of the partial paths for each vehicle. Algorithm 2 describes the operation of each vehicle in this scheme.

Algorithm 2 RSU(**RSUAddr**, **PT**[ ][ ])

```
beaconTimer = X;
start beaconTimer;
while true do
    if beaconTimer == 0 then
        restart beaconTimer;
        broadcast (RSUBeacon); //which contains RSUAddr
        receive (V_i and PP_i[ ]) from vehicles;
        if new V_i then
            create new entry forV_i in pathsTable PT[ ][ ];
            append PP_i[ ] to PT[V_i][ ];
        end if
    else
        append PP_i[ ] to PT[V_i][ ];
    end if
end while
```

As can be observed from Algorithm 2, each RSU will maintain a broadcast timer *beaconTimer* which will trigger the *RSUBeacon* message to be broadcasted to all vehicles in its vicinity. In addition, the *beaconTimer* is reset after every timer expiration for the next broadcast. Each RSU has an *RSUAddr* which identifies it and paths table PT which stores all the received partial paths PP_i from the vehicles in its transmission range. The PT here will append the partial path information to a vehicle's V_i record if it already exists, or will create a new entry for a newly discovered vehicle and then append the partial path received.

This is the only scheme in the proposed work that requires RSUs to query the vehicles for their collected paths. Vehicles will continue collecting the road segments S_i and appending them to their local partial paths PP_i. Once a beacon message from a neighboring RSU is received, they will provide it with the collected PP_i.

4.2 Vehicle-Triggered Schemes - Broadcast Mode

In the Vehicle-triggered schemes to come, the communication is triggered by events occuring at the vehicles' end. The RSUs are only responsible for receving and storing the collected path from the vehicles in their vicinity. This class of schemes runs under two modes: Broadcast Mode and RSU Discovery Mode. In the Broadcast mode, vehicles send their collected path information in a broadcast manner to the surrounding RSUs. In the RSU Discovery Mode however, vehicles locate the closest RSU and then send their path information to that RSU. In the next subsections, two Vehicle-triggered schemes running under the Broadcast Mode will be explained in details alongwith an algorithm to describe the operation of each vehicle and RSU in a clear and systematic manner.

4.2.1 Newest-Edge (VTB-NE)

For this scheme, vehicles will also be responsible for collecting the road segments S_i as they traverse them, but the RSU database communication updates will only occur every time a vehicle collects a new road segment. In this mode, vehicles will travel through the VANET and record every new S_i they encounter during their trip. Once this new S_i is collected, the vehicle will create a packet containing the attributes shown in Figure 4.3. As can be noticed from the last attribute, only one road segment (i.e. newEdge in Figure 4.3) is sent to the neighboring RSU(s). The packet is sent in a broadcast fashion to all RSUs in the vicinity, which may allow for more than one receipt of the new road segment. Algorithm 3 describes the operation of each vehicle in this scheme.

Algorithm 3 Vehicle($\mathbf{V}_i$, $\mathbf{S}_i$)

```
while true do
    get road segment S_i;
    if new S_i then
        broadcast (V_i and S_i) to all neighboring RSUs;
    end if
end while
```

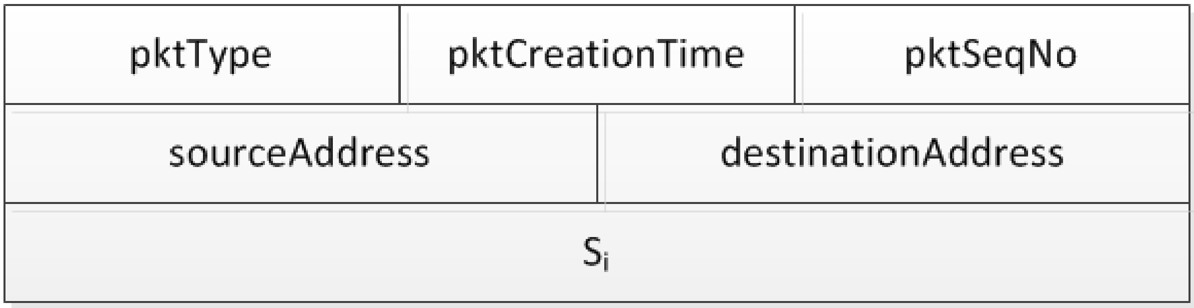

Figure 4.3: Packet design for Newest-Edge Vehicle-Triggered Scheme

For RSUs, messages containing the new S_i collected by the vehicles will be received and appended to their database accordingly. If the received message is from a new vehicle, the vehicle ID will be added to the database, including the road segment S_i in the received message. If, however, the message is received from a vehicle with an already existing entry in the RSU's database, the new road segment in the message is appended to the existing vehicle ID's path. Similar the RSU-triggered scheme described previously and the rest of the communication schemes to come, the pktCreationTime attribute can be used to sort the received road segments accordingly to form a timely-ordered sequence of events (i.e., detection of a new road segment). Also, the RSUs can be triggered at the end of the collection period to send their stored vehicle path information to a central server. Algorithm 4 describes the operation of each RSU in this scheme.

Algorithm 4 RSU(**RSUAddr**, **PT**[ ][ ])

while true **do**
    if ($\mathbf{V}_i$ and $\mathbf{S}_i$) is received **then**
        if new $\mathbf{V}_i$ **then**
            create new entry for$\mathbf{V}_i$ in pathsTable **PT**[ ][ ];
            append $\mathbf{S}_i$ to $\mathbf{PT}[\mathbf{V}_i]$[ ];
        end if
    else
        append $\mathbf{S}_i$ to $\mathbf{PT}[\mathbf{V}_i]$[ ];
    end if
end while

Since the RSU does not need to periodically trigger vehicles to send their information, no beacon timer was required in this scheme. The RSU will receive all new road segments S_i from all vehicles in its transmission range and add them accordingly to its paths table PT as mentioned previously.

4.2.2 Full Path (VTB-FP)

Similar to the two vehicle-triggered schemes described above, vehicles in this scheme are also solely responsible in deciding when to send information to the RSU. In this scheme, the vehicles keep track of all the road segments S_i as they travel through the VANet in order of arrival to those segments. However, the vehicles will only broadcast their full path information $\boldsymbol{FP}_i$ to the surrounding RSUs once they have reached their final destination (i.e. once the vehicle has stopped moving). Figure 4.4 shows the contents of the packet sent from the vehicle to the RSUs in its vicinity. This packet layout is also used for the alternative communication mode of this same scheme (i.e., RSU Discovery Mode). Algorithm 5 shows the operation of each vehicle in this scheme.

Algorithm 5 Vehicle($\mathbf{V}_i$,$\mathbf{FP}_i$[ ])

```
while true do
    if V_i is moving then
        get road segment S_i;
        if new S_i then
            append S_i to full path FP_i[ ];
        end if
    end if
    if V_i has reached final destination then
        broadcast (V_i, FP_i[ ]) to all neighboring RSUs;
    end if
end while
```

Once received by the RSU, the RSU will add this path information to its database. Similar to Newest-Edge scheme, the RSU does not need to periodically trigger vehicles

<table>
<tr><td colspan="2">pktType</td><td colspan="2">pktCreationTime</td><td colspan="2">pktSeqNo</td></tr>
<tr><td colspan="3">sourceAddress</td><td colspan="3">destinationAddress</td></tr>
<tr><td colspan="2">pathSize</td><td colspan="4">FP_i</td></tr>
</table>

Figure 4.4: Packet design for Full Path Vehicle-Triggered Scheme

to send their information, and therefore no beacon timer was required in this scheme. Algorithm 6 shows the operation of each RSU in this scheme.

Algorithm 6 RSU(**RSUAddr**, **PT**[ ][ ])

while true **do**
  if ($\mathbf{V}_i$, $\mathbf{FP}_i$[ ]) is received **then**
    create new entry for$\mathbf{V}_i$ in pathsTable **PT**[ ][ ];
    append $\mathbf{FP}_i$[ ] to $\mathbf{PT}[\mathbf{V}_i]$[ ];
  end if
end while

The RSU will receive all the full paths FP_i from all vehicles in its transmission range and add them to its paths table PT as they arrive. There will also no need to check if a certain vehicle's ID, associated to the full path sent, already exists since the path is complete and there is no need to append to any existing entries.

4.3 Vehicle-Triggered Schemes - RSU Discovery Mode

A modification to the previous Vehicle-triggered schemes was introduced in this section in order to send the newest collected path information to the closest RSU in a unicast manner. Only after a vehicle locates the closest RSU will it send its collected path information. In this section, both Vehicle-triggered schemes mentioned in the previous section will be described in details with regards to the new mode they will be running in (i.e., RSU Discovery Mode), and an algorithmm describing the operation of each vehicle and RSU will be provided.

4.3.1 Newest-Edge (VTRD-NE)

In this scheme, the vehicles collect the information about the road segments S_i they are currently on. Once a new segment is entered, the vehicle will first find the closest RSU by broadcasting an *RSUDiscovery* message which will request RSUs in its vicinity to send back their addresses. According to [41], the Time of Arrival (ToA) is a decent indication of the physical location of a certain node. In that case, the vehicle should only accept the first *RSUAddr* message it receives, since it will be the closest one due to the time of arrival for that message to the requesting vehicle is the shortest. Once the address of the closest RSU is received, the vehicle will then send the generated packet containing the newest road segment S_i entered, including its vehicle ID. The packet sent from the vehicle with the newest road segment S_i detected has the same layout design as that in Figure 4.3. Algorithm 7 shows the operation of each vehicle in this scheme.

Algorithm 7 Vehicle($\mathbf{V}_i$, $\mathbf{S}_i$)

```
while true do
    get road segment S_i;
    if new S_i then
        broadcast (V_i, RSUDiscoveryBeacon) to all neighboring RSUs;
    end if
    if (RSUAddr) received then
        send (V_i, S_i) to RSUAddr;
    end if
end while
```

Upon receipt of the *RSUDiscovery* message, RSUs will send back an *RSUAddr* message containing its address to the requesting vehicle. For some vehicles, more than one RSU may receive the *RSUDiscovery* message depending on their location, which will cause them all to reply to the requesting vehicle. Algorithm 8 shows the operation of each RSU in this scheme.

Once a vehicle sends its new road segment S_i to the closest RSU, that RSU will append

Algorithm 8 RSU(**RSUAddr**, **PT**[ ][ ])

```
while true do
    if (V_i and RSUDiscoveryBeacon) is received then
        send (RSUAddr) to V_i;
    end if
    if (V_i and S_i) is received then
        if new (V_i then
            create new entry forV_i in pathsTable PT[ ][ ];
            append S_i to PT[V_i][ ];
        end if
    else
        append S_i to PT[V_i][ ];
    end if
end while
```

this new road segment to its path table PT accordingly. Similar to the operation of the Newest-Edge Broadcast Mode scheme, if the received message is from a new vehicle, the vehicle ID will be added to the database, including the road segment S_i in the received message. If, however, the message is received from a vehicle with an already existing entry in the RSUs database, the new road segment in the message is appended to the existing vehicle IDs path.

4.3.2 Full Path (VTRD-FP)

Following the same reasoning in the Newest Edge Scheme - RSU Discovery Mode, this scheme will perform an RSU discovery before sending the full working path information FP_i to the closest RSU. When a vehicle reaches its final destination, it first broadcasts an *RSUDiscovery* message to all surrounding RSUs, querying for their addresses. Once the vehicle determines the address of the closest RSU, it will send the full path FP_i collected during its trip to that RSU. The packet sent from the vehicle with the full path FP_i collected has the same layout design as that in Figure 4.4. Algorithm 9 shows the operation of each vehicle in this scheme.

Upon receipt of the *RSUDiscovery* message, RSUs will send back an *RSUAddr* mes-

Algorithm 9 Vehicle($\mathbf{V}_i$, $\mathbf{S}_i$)

while true **do**
  if $\mathbf{V}_i$ is moving **then**
    get road segment $\mathbf{S}_i$;
    if new $\mathbf{S}_i$ **then**
      append $\mathbf{S}_i$ to full path $\mathbf{FP}_i$[ ];
    end if
  end if
  if $\mathbf{V}_i$ has reached final destination **then**
    broadcast ($\mathbf{V}_i$, **RSUDiscoveryBeacon**) to all neighboring RSUs;
    if (**RSUAddr**) received **then**
      send ($\mathbf{V}_i$, $\mathbf{FP}_i$[ ]) to **RSUAddr**;
    end if
  end if
end while

sage containing its address to the requesting vehicle. For some vehicles, more than one RSU may receive the RSUDiscovery message depending on their location, which will cause them all to reply to the requesting vehicle. Algorithm 10 shows the operation of each RSU in this scheme.

Algorithm 10 RSU(**RSUAddr**, **PT**[ ][ ])

while true **do**
  if ($\mathbf{V}_i$ and **RSUDiscoveryBeacon**) is received **then**
    send (**RSUAddr**) to $\mathbf{V}_i$;
    if ($\mathbf{V}_i$, $\mathbf{FP}_i$[ ]) is received **then**
      create new entry for$\mathbf{V}_i$ in pathsTable **PT**[ ][ ];
      append $\mathbf{FP}_i$[ ] to **PT**[$\mathbf{V}_i$][ ];
    end if
  end if
end while

4.4 Summary

In this chapter, five novel communication schemes have been designed to prepare vehicular movement paths for sequential patterns data mining. The communication that occurs is between vehicles and RSUs on the road, which is, as described in Chapter 1, is a vehicle-to-infrastructure (V2I) type of communication. Each scheme was described and an algorithm followed by each vehicle and RSU for each of those schemes has been provided. In the next chapter, these schemes will be evaluated against network performance metrics to compare them with each other and evaluate their efficiency in terms of those metrics. The movement patterns collected from these schemes will also be data mined in order to observe the effects of sequential data mining on the collected data.

Chapter 5

Performance Evaluation

In this chapter, both the data preparation communication schemes and the sequential data mining aspects of this work have been investigated. The Vehicular Movement Patterns Data Preparation Schemes have been developed and implemented in the Network Simulator 2 (NS-2) [42] environment to simulate a realistic environment. The map layout used in this performance evaluation is a Manhattan 2-D grid layout, as shown in Figure 5.1. Performance evaluation metrics were captured for each scheme and an analysis was performed to present their strengths and weaknesses. In addition, an analysis of the data mining aspect of this work has also been discussed, showing the effect of a user-defined support threshold on the number of frequent vehicular movement patterns that can be extracted from the collected vehicular paths and the confidence associated with those patterns. For the sake of simplicity, each scheme described in Chapter 4 has been given a shorter term to represent it, and Table 5.1 will represent this change.

Scheme	**Name**
RT	RSU-Triggered Scheme
VTB-NE	Vehicle-Triggered Broadcast Mode - Newest Edge
VTB-FP	Vehicle-Triggered Broadcast Mode - Full Path
VTRD-NE	Vehicle-Triggered RSU Discovery Mode - Newest Edge
VTRD-FP	Vehicle-Triggered RSU Discovery Mode - Full Path

Table 5.1: Evaluated Communication Schemes

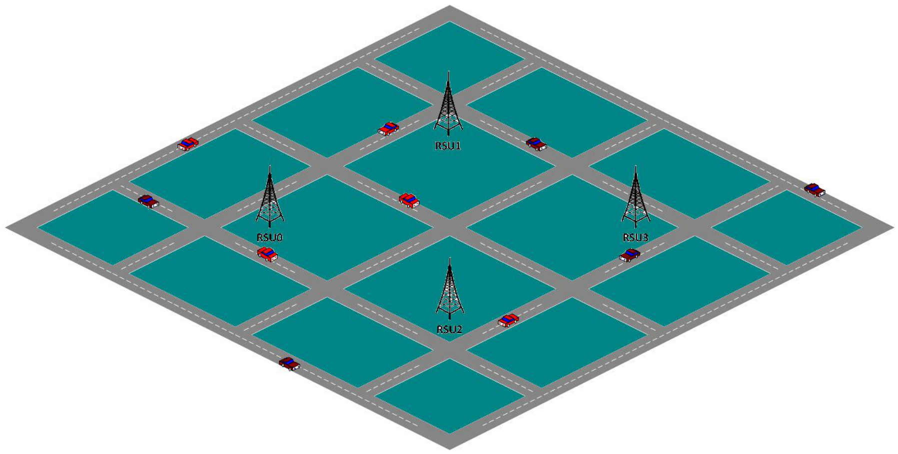

Figure 5.1: Grid Map with 4 RSUs

5.1 Simulation and Analysis Parameters, Process and Metrics

In the implementation of the proposed communication schemes, all vehicles and RSUs were set to operate with certain specific parameters. Table 5.2 shows the general simulation parameters used in the implementation and evaluation of the proposed schemes.

Each RSU was positioned in a fixed location on the map shown in Figure 5.1 shown previously. The location of the RSU was chosen in a way that allows for their communication range to cover the whole area of the map which will encompass all vehicles traveling along those road segments in the same given map. Vehicles were given mobility paths to follow that were set either manually or randomly depending on the performance metric to be measured. As the vehicles travel through the given map, they perform a map matching process that allows them to locate which road segment they are currently on and add this road segment to their collected path information. The map matching process was also included in the implementation, which translated the Cartesian coordinates vehicles were located on into a unique road segment identifier.

In order to quantitatively evaluate the proposed communication schemes, different network performance metrics were collected through an extensive set of experiments.

Parameter	**Value**
Channel Type	Wireless Channel
Radio Propagation Model	Two-Ray Ground
MAC Type	IEEE802.11p
Communication Range (m)	300
Map Layout	Manhattan Grid 4x4
Map Area (m^2)	640 000
Number of Road and Junction Segments	105
Number of Vehicles	10 - 70
Vehicular Speed (m/s)	14
Trip Path Size	Dependent on Experiment
Map Matching Interval (s)	1
Number of RSUs	4
RSU Beacon Broadcast Interval (s)	10
Simulation Time (s)	800

Table 5.2: Schemes Simulation Parameters

These metrics are:

- Communication Overhead
- Average Packet Delay
- Average Packet Delivery Ratio

For the data mining evaluation, frequent vehicular movement patterns were extracted from the sample of collected vehicular paths in one of the schemes. As discussed in Chapter 3, the user-defined minimum support threshold and confidence of generated movement rules are the two important factors that can affect the extraction of frequent sequential patterns and probability that a generated rule will occur respectively. Since the paths collected from the vehicles during the collection process are in the form of timely-ordered events, as sequential patterns should be, the effect of these two factors on the number of frequent patterns extracted and prediction are analyzed. The parameters shown in Table 5.3 provide the analysis parameters used in this analysis.

Parameter	Value
Vehicular Paths Sample Size	225
Minimum Support Threshold	10% to 50%

Table 5.3: Data Mining Analysis Parameters

5.2 Performance Evaluation of the Proposed Data Preparation Schemes

In this section, the proposed implemented data preparation schemes are evaluated with respect to the impact they impose on the network performance. Each performance metric is defined, the results associated with those metrics are shown and a discussion explaining those results is provided in this section.

5.2.1 Communication Cost

In this set of extensive experiments, the communication overhead was evaluated for each scheme. The communication overhead in this context is the number of messages sent during the simulation run time by all participating entities in the given map (i.e., RSUs and vehicles). This communication overhead was measured with respect to (1) the *number of segments* traversed by each vehicle and (2) the *number of vehicles* in the given map. For the first variable (i.e., number of segments), a number of vehicles were set to traverse a path with an increasing number of segments and increasing number of vehicles. The number of messages exchanged during the simulation run was collected and is considered the communication overhead associated to the path size. Since our data collection schemes are responsible mainly for collecting vehicular paths, the number of segments in those paths will vary and there is a need to evaluate the efficiency of our schemes in regards to that. For the second variable, the aim was to measure the same overhead incurred with only the number of vehicles varying with randomly generated path sizes.

Overhead with respect to Path Size

In these experiments, the number of vehicles was set to increase from 10 to 70 with increments of 10 vehicles every simulation run and the path size was increased from 5 to 20 traversed segments with increments of 5 segments for each simulation run. The number of messages sent from the traveling vehicle to the RSUs and vice-versa in the given map was included in the measurement. Figure 5.2 shows the overhead incurred by the implementation of each scheme in the aforementioned grid map with respect to the different path sizes.

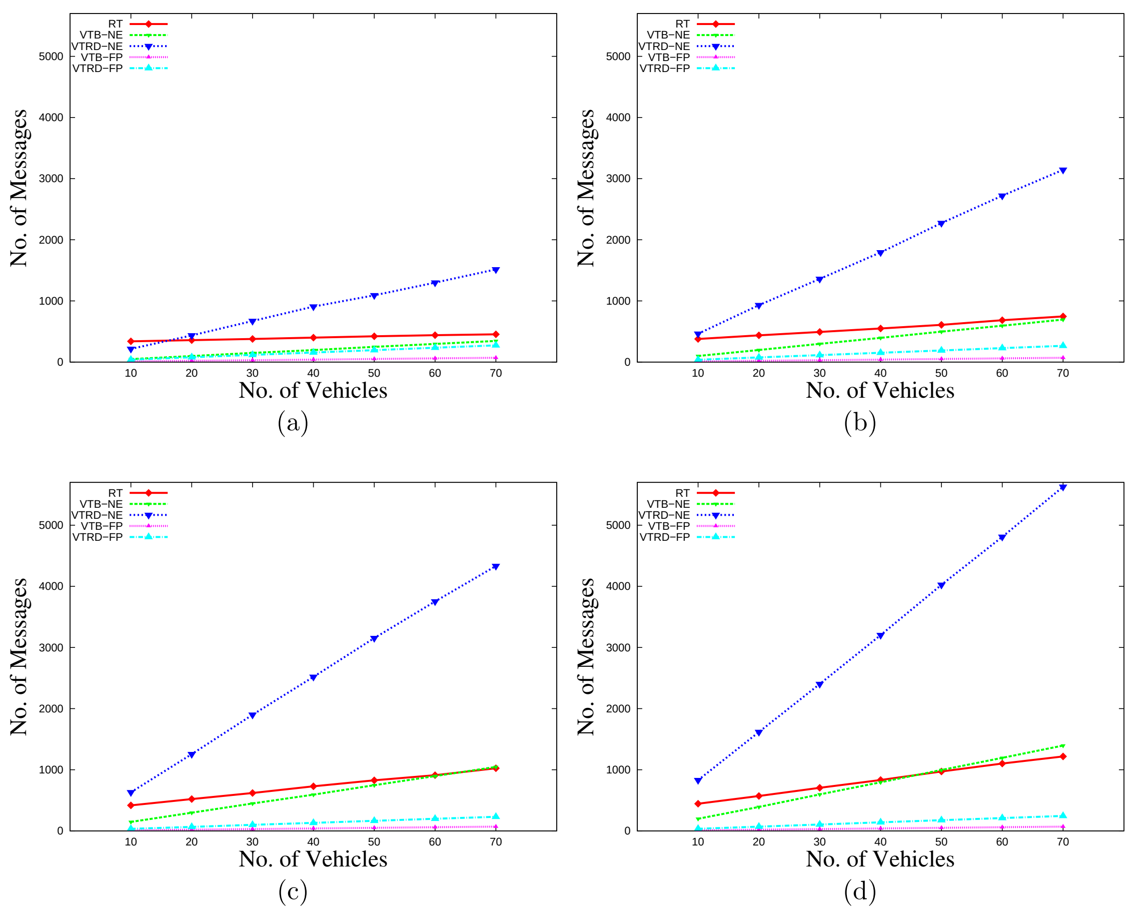

Figure 5.2: Network Communication Overhead with respect to Path Size **(a)** Path Size = 5 segments, **(b)** Path Size = 10 segments, **(c)** Path Size = 15 segments, and **(d)** Path Size = 20 segments

As can be observed from the results illustrated in Figure 5.2, it is apparent that VTRD-NE is the most costly in terms of communication overhead. The main reason for this lies in the fact that for every new road segment detected by each vehicle, an RSU Discovery needs to be performed. Compared with VTB-NE, the results of VTRD-NE show a communication overhead that is almost 400% more, even though the number of messages exchanged between a specific vehicle and RSU in VTRD-NE should only increase the communication by 300% due to 3 messsages sent (i.e., the *RSUDiscovery* message, the *RSUAddr* reply message and the *NewEdge* message). This is explained by the receipt of the *RSUDiscovery* message by more than one RSU in the communication range of the vehicle. Some of the vehicles, when a new road segment was detected, were in the vicinity of more than one RSU, and each recipient of the *RSUDiscovery* message would have to reply, which may cause the number of *RSUAddr* replies received by a vehicle to double or even triple based on the number of RSUs replying.

RT's communication overhead showed very interesting results compared to the other schemes. As the number of road segments increased, so did the number of messages sent by the vehicles. However, the number of messages sent by RT also include the beacon message sent by each RSU every 10 seconds. This beacon message is sent regardless of the presence of vehicles in the communication range of any RSU.

VTB-FP and VTRD-FP showed the most promising results in terms of communication overhead, since the vehicles will only send their full final path when they have reached their final destination. For VTB-FP, only one message was sent to the surrounding RSU by means of broadcasting, regardless of the number of paths traversed by each vehicle. VTRD-FP however required each vehicle to first discover a neighbouring RSU and then send its full collected path to that discovered RSU. Similar to VTB-NE and VTRD-NE, the number of messages sent during the operation of VTRD-FP is approximately 4 times higher than that of VTB-FP. VTB-FP and VTRD-FP significantly improved the communication overhead compared to RT, VTB-NE and VTRD-NE.

Overhead with respect to Number of Vehicles

In this set of experiments, the communication overhead was measured with respect to the number of traveling vehicles in the same map shown in Figure 5.1. As explained before, the number of vehicles was varied starting from 10 vehicles to 70 vehicles, with an increment of 10 vehicles every simulation run. The path size in this experiment was not specified and was generated randomly to mimic a realistic setting where vehicles travel in a given map according to their intended destination. The experiments were performed for a time period of 800 seconds and the number of messages exchanged between the vehicles and their surrounding RSUs was collected. Figure 5.3 shows the results of the experiments.

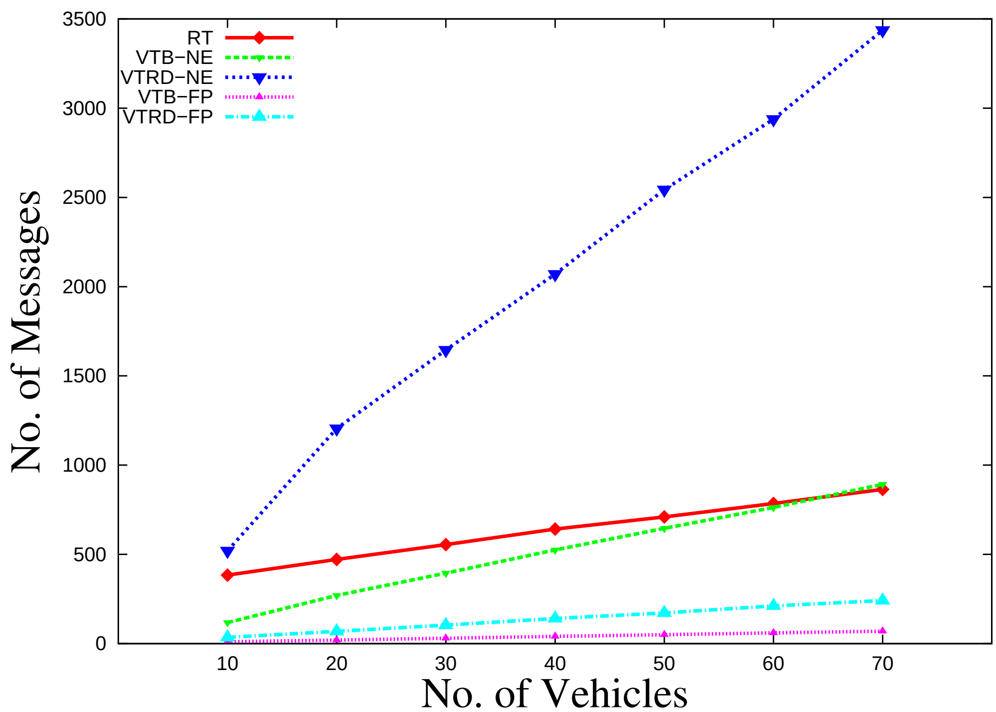

Figure 5.3: Network Communication Overhead Analysis with respect to Vehicular Density

VTB-NE and VTRD-NE in this set of experiments show results that are very interesting. The number of messages exchanged in VTRD-NE is more than 3 times higher than the number of messages exchanged in VTB-NE. After analyzing the exchange of messages more closely, it was observed that the reason for this higher increase rate was due to a number of vehicles receiving the RSUAddr reply message from more than one RSU during their RSU discovery phase. Many of the vehicles, when a new road segment is entered, were in the communication range of 2 RSUs, and these RSUs had to each send an RSUAddr reply message. In VTB-NE however, only the vehicles will broadcast their newly discovered road segment to the neighbouring RSUs during their trip.

In VTB-FP and VTRD-FP, the least amount of overhead was noticed, since the vehicles only attempt to update the RSUs with their full paths once they have reached their final destination. VTRD-FP shows a higher number of messages exchanged between the vehicles and RSUs than VTB-FP. This is explained by the need for VTRD-FP, as previously mentioned, to discover the RSU before sending its full path. Similar to VTB-NE and VTRD-NE, the increase rate factor is more than 3 here as well due to the increased number of replies to some of the vehicles' *RSUDiscovery* messages caused by overlapping communication ranges between two or more neighbouring RSUs.

5.2.2 Latency

Since the aim of these experiments is to evaluate the performance of the proposed schemes, the average packet delay is an important efficiency metric that needs to be measured. In this set of experiments, the average packet delay is defined as the time period between the creation of the message containing the vehicular path information at the vehicle's side and its successful receipt at the respective RSU. This definition is derived and extended from the definition of packet delay in [43]. For the schemes proposed, an obvious difference in the average packet delay should be noticed for schemes that may require more time in order to prepare their collected path information for sending. The significance of this metric lies in the type of application these schemes will serve. For instance, real-time operations that require prompt path information may be more sensitive to average packet delay as opposed to other less delay-sensitive applications.

In this set of experiments, the average packet delay is observed with respect to the number of vehicles in the network. Similar to the communication overhead analysis, the number of vehicles was increased from 10 vehicles to 70 vehicles in increments of 10 vehicles per simulation run. In this evaluation, the collected results were grouped in a manner that displays the average packet delay for each group of schemes with similar packet sizes. More specifically, the average packet delay for RT is displayed individually in Figure 5.4(a), results for VTB-NE and VTRD-NE are displayed in Figure 5.4(b) and VTB-FP and VTRD-FP are shown in Figure 5.4(c). For each scheme, the average packet delay is computed at the RSUs once the path information has been received. In order to do so, each RSU makes use of the *packetCreationTime* attribute available in the packet received, which contains the time at which the packet was initially created. These individual delays are then collected and averaged over the number of successfully received messages.

Before discussing the results shown in Figure 5.4, there is a need to understand the different sources of delay associated to the operation of the schemes proposed. The following are the main sources of delay incurred by the schemes:

- **Creation and Processing** - The creation delay is the amount of time required by each node to generate a packet and prepare it for delivery. The processing delay refers to the time required by each node to process and extract the required information from the received packet.

- **Propagation** - The propagation delay is the amount of time taken for the packet to reach its destination after transmission.[49] [50]

- **Queueing** - The queueing delay is the time a packet spends in each nodes buffer waiting to be processed. [47] [48]

- **TX/RX Switching** - The TX/RX switching delay is the time required by a node to switch from the transmitting channel (TX) to the receiving channel (RX), and vice-versa. [45] [46]

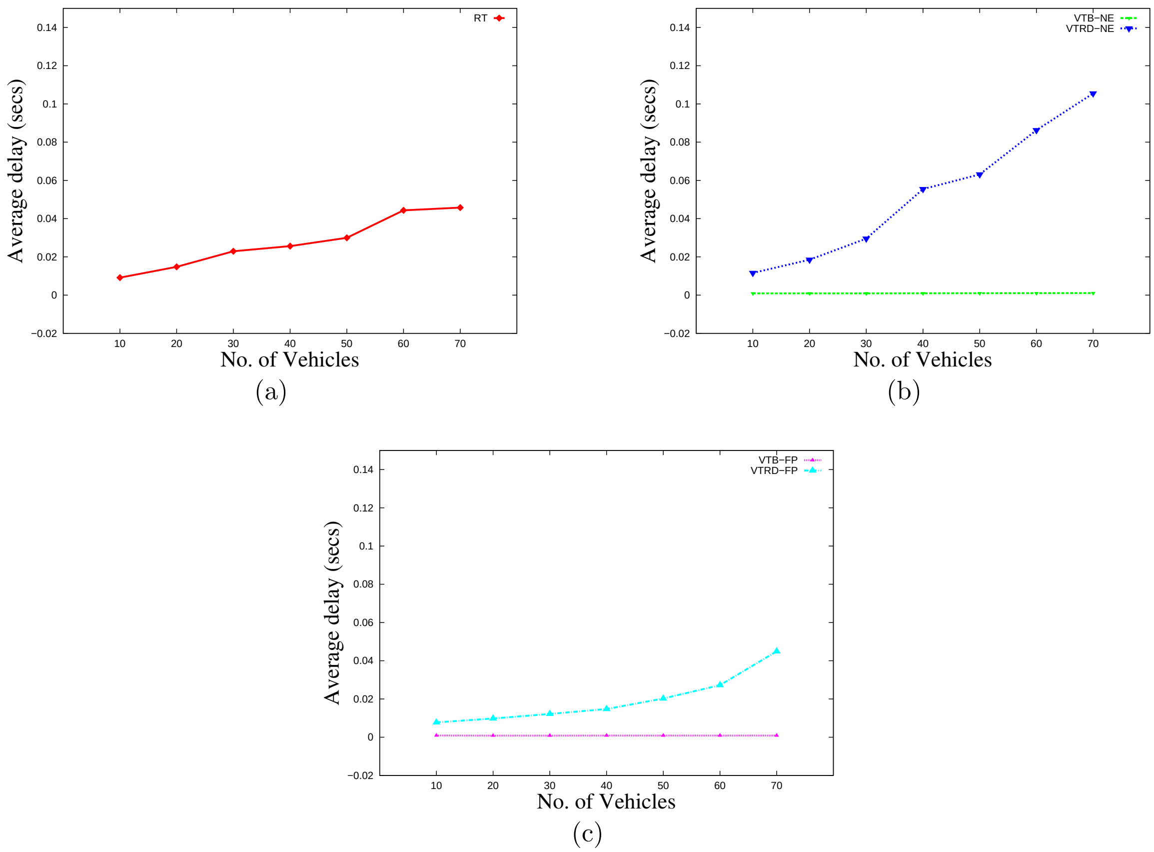

Figure 5.4: Average Packet Delay for (a) RT, (2) VTB-NE and VTRD-NE, and (3) VTB-FP and VTRD-FP

Taking into consideration these delay sources, the results in Figure 5.4 reveal the difference in average packet delay achieved by each proposed scheme. From the analysis performed, it is obvious that RT, VTRD-NE and VTRD-FP have introduced higher average packet delay than VTB-NE and VTB-FP. For VTRD-NE and VTRD-FP, the reason behind this difference lies mainly in the delay incurred by the RSU discovery phase. In both these schemes, vehicles will create their packets when, for VTRD-NE, they discover a new road segment or, for VTRD-FP, they reach their final destination. In VTB-NE and VTB-FP however, the packet is created and sent as soon as a new road segment is detected or the vehicle has reached its final destination respectively. RT's increasing average packet delay, however, is caused by the increased queuing that occurs at the RSU's side when receiving multiple packets from the vehicles queried for their partial path information. Since vehicles will receive the broadcasted beacon at approximately the same time, the RSUs will have to collect numerous replies from the vehicles in their vicinity, increasing the queuing delay. Also affected by queuing delay is the significantly high average packet delay in VTRD-NE, mainly due to the frequency at which vehicles will need to send their packets which increases this queuing delay.

VTB-NE and VTB-FP show promising results in terms of average packet delay. This is mainly due to the fact that packets are broadcasted immediately when required to do so, as per their operation, which alleviates the need for idle waiting during the RSU discovery phase in VTRD-NE and VTRD-FP.

In general, RT, VTRD-NE and VTRD-FP are also higher due to the high number of switching between TX (transmission) and RX (receiving) modes. VTRD-NE is significantly higher also for this reason, since it forces both the vehicles and RSUs to switch between both modes at every new road segment.

5.2.3 Effectiveness and Reliability

In order to evaluate the reliability and effectiveness of the proposed scheme, the packet delivery ratio was measured. The reliability of the proposed schemes is very important, as the sequential patterns need to be formed accurately in order to be used in the data mining and prediction phases. In these experiments, the packet delivery ratio is the ratio between the number of packets originating from the vehicles and the number of successfully received packets at the RSU [44]. The packet delivery ratio was measured only for those packets containing path information to be sent from the vehicles. For each scheme, an extra attribute named sequence number, pktSeqNo shown in Figures 4.2, 4.3, and 4.4, was added to the packets being sent from the vehicles in order to keep track of the sequence number of all the messages sent by the vehicles and received by the RSUs. Each experiment was set to calculate the packet delivery ratio of each scheme using the following equation:

$$PDR_{TOTAL} = \frac{\Sigma PDR_{PACKET}}{No.of SentMessages} \tag{5.1}$$

where,

$$PDR_{PACKET} = \frac{RCVDMsg_{pktSeqNo}}{SENTMsg_{pktSeqNo}} \tag{5.2}$$

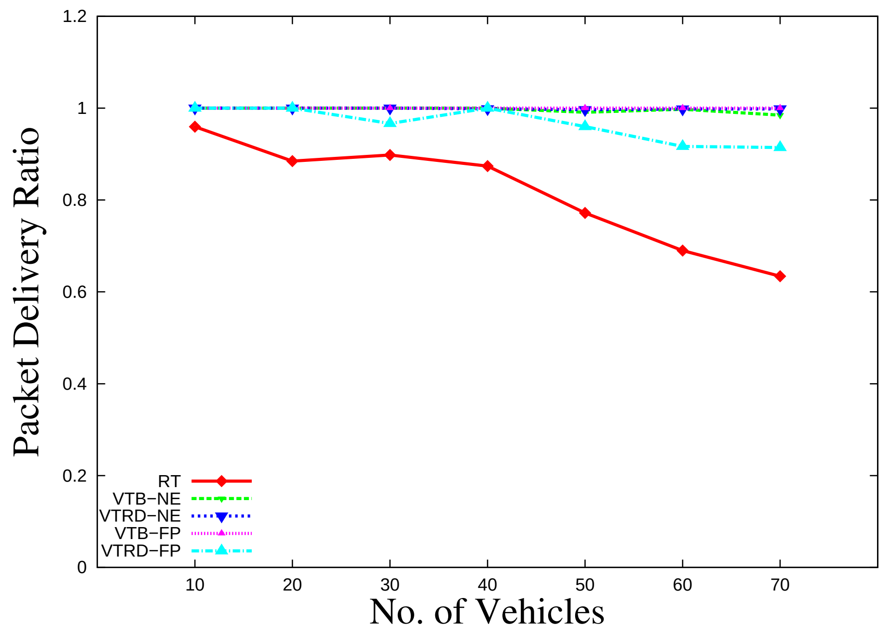

Figure 5.5: Average Packet Delivery Ratio with respect to Vehicular Density

As can be observed from the results shown in Figure 5.5, RT's packet delivery ratio seems to suffer when the number of vehicles increases. Similar to the significant difference in average packet delay, the reason behind this downfall is the time period in which all vehicles need to send their partial path information. Since vehicles neighboring an RSU will receive the *RSUBeacon* message at approximately the same time, they will also send their messages almost simultaneously. This will cause the vehicles to send their partial path information to the neighbouring RSU at relatively the same time as other vehicles in the RSUs vicinity. With the increase in the number of vehicles sending their messages at once, the number of packet collisions increases, causing some of the packets to be dropped instead of arriving at their destination successfully.

VTB-NE, VTRD-NE, VTB-FP and VTRD-FP suffer almost no packet losses due to the timing of their path information update communication being independent of any synchronized transmission schedule. VTB-NE and VTB-FP achieved a packet delivery ratio PDR_{packet} of greater than 1 for some packets due to their broadcasted path information being received at more than one RSU. This case occurs when a vehicle, when ready to send its information, is in the communication range of more than one RSU.

5.2.4 Performance Evaluation Summary

In the previous subsections, the efficiency and effectiveness of the proposed data collection schemes were evaluated. The performance metrics used for this evaluation were communication overhead, average packet delay and average packet delivery ratio.

From the communication overhead analysis, with regards to the size of the path traversed with an increasing number of vehicles, it can be concluded that VTB-FP and VTRD-FP incur a lower communication overhead than the other proposed schemes. Further analysis also showed that VTB-FP and VTRD-FP introduce a lower overhead compared to the other schemes when the vehicle density in the given map increases. In scenarios where communication overhead is an issue, RT showed a considerable amount of overhead due to periodic transmission of the *RSUBeacon* message, but still performed better than VTRD-NE. VTRD-NE may not be an appropriate option due to the increasing amount of overhead incurred on the network with respect to the path size and vehicle density.

In terms of average packet delay, VTRD-NE and VTRD-FP showed the least amount of delay incurred. The delay may not be too important during the ***Data Gathering*** phase, but might be critical during the ***Prediction and Monitoring*** phase where real-time updates may be required in order to predict the path of a targeted vehicle as it is traveling in a geographical map.

The reliability of the proposed schemes was evaluated using the packet delivery ratio.

From the results shown in Figure 5.5, it was apparent that RT is not very tolerant of increased vehicle density. RT may be more suitable to lesser dense settings such as highways and suburban areas, where the probability of packet collision will be lower, allowing for a higher packet delivery ratio. The vehicle-triggered schemes (i.e., VTB-NE, VTRD-NE, VTB-FP and VTRD-FP) also suffered almost no packet losses.

Table 5.4 shows a summary of the performance evaluation done on the five collection schemes. Each scheme is listed in the table, the metrics are compared and the more suitable phase to each scheme is outlined.

	Performance Comparison			**Suitable Phase**	
Scheme	*Communication Overhead*	*Average Packet Delay*	*Average Delivery Ratio*	*Data Gathering*	*Prediction and Monitoring*
RT	High	Moderate	Low	**X**	✓
VTB-NE	Moderate	Moderate	High	**X**	✓
VTRD-NE	High	High	High	**X**	**X**
VTB-FP	Low	Low	High	✓	✓
VTRD-FP	Low	Low	High	✓	✓

Table 5.4: Data Preparation Schemes Comparison

5.3 Analysis of Vehicular Movement Patterns Data Mining

In this section, the collected vehicular paths by one of the schemes are data mined in order to extract the most frequent movement patterns and the confidence for generated movement rules is obtained. In order to extract frequent movement patterns, a user-defined minimum support threshold needs to be set. This minimum threshold will signify the sensitivity of the extraction process, since it will affect the number of movement patterns that can be considered as frequent. As for the confidence, generated movement rules will be generated for some frequent movement patterns extracted. This section will cover the analysis of sequential patterns data mining with respect to two important factors; minimum support and confidence.

5.3.1 Frequent Movement Patterns Extraction

In order to extract frequent movement patterns from the sample of vehicular paths collected, a minimum support threshold is set by the user. In this analysis, a set of support threshold have been used to study their effect on the extraction process, ranging from 10% to 50%. A sample of 225 collected vehicular paths participated in this analysis, and the number of extracted frequent patterns with respect to those defined support thresholds was reported. For instance, if a minimum support of 10%, all vehicular sub-patterns occuring in the sample of collected vehicular paths with a support of 10% or higher were calculated and their total frequency was recorded. Since, as explained in Chapter 3, the minimum support threshold sets the sensitivity of extraction process, the number of frequent patterns extracted should be inversely affected by an increase in the defined minimum support. Figure 5.6 shows the result of the data mining performed on the sample of collected vehicular paths.

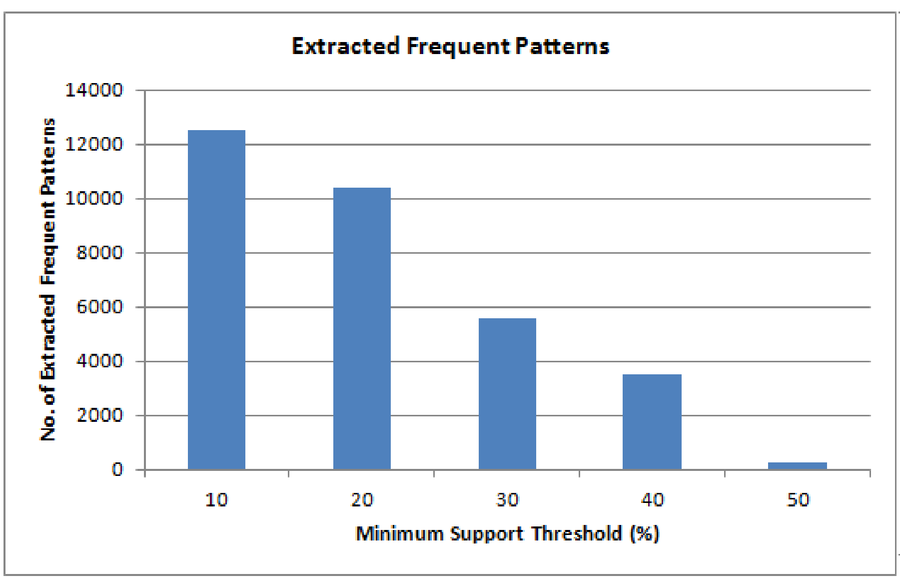

Figure 5.6: No. of Frequent Vehicular Movement Patterns Extracted

From the results shown in Figure 5.6, it is apparent that the number of vehicular subpatterns that can be considered as being frequent in the collected vehicular paths sample decreases as the minimum support threshold increases. This can be explained by the fact that many subpatterns have a support that is higher than the set minimum support threshold initially at 10%, and can be classified as frequent subpatterns for that set threshold. From the figure, the number of times each vehicular subpattern has occured as a frequent one in the sample of collected vehicular paths was approximately 12500 when the minimum support was set to 10%. When at 50%, the minimum support threshold has lowered the number of frequent subpatterns to almost 500, which shows a decrease of almost 96% in the number of extracted patterns. More specifically, the percentage of decrease in the number of extracted frequent subpatterns between a minimum support of 10% and 20% is 16%, between 20% and 30% is 47.6%, between 30% and 40% is 36.4% and between 40% and 50% showed the highest decrease of almost 86%.

5.3.2 Confidence of Generated Movement Rules

In this analysis, the confidence of generated movement rules was investigated. The confidence, as discussed in Chapter 3, refers to the probability of different generated movement rules from the frequent movement patterns extracted. For the sake of this analysis, a junction shown in Figure 5.7 and the different subpatterns around it were used to generated movement rules.

In order to generate all the possible movement rules for vehicles coming from road segments S_{58}, S_{112}, the collected vehicular paths were scanned thoroughly for subpatterns where S_{58}, S_{112} occured in that same sequence. From the analysis, all the possible movement rules generated are shown below:

- **Rule 1** : S_{58} , $S_{112} \Rightarrow S_{112}$ (i.e., stays at intersection)
- **Rule 2** : S_{58} , $S_{112} \Rightarrow S_{19}$ (i.e. turns left)
- **Rule 3** : S_{58} , $S_{112} \Rightarrow S_{60}$ (i.e. continues straight ahead)

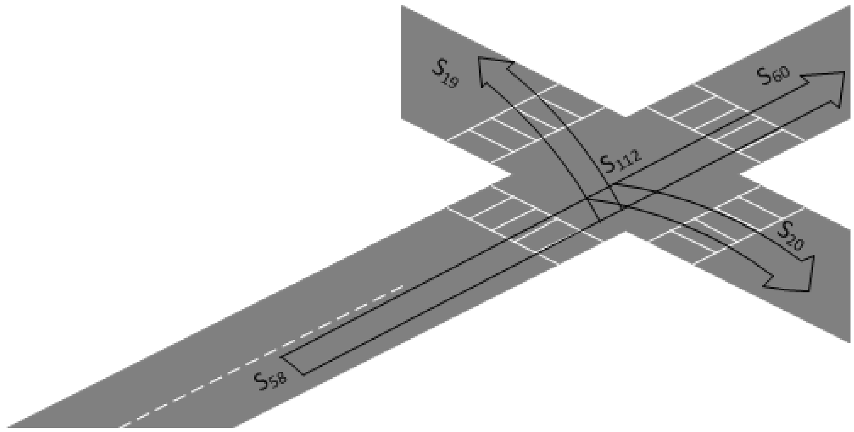

Figure 5.7: Movement Rules for Subpattern S_{58}, S_{112}

- **Rule 4** : S_{58} , $S_{112} \Rightarrow S_{20}$ (i.e. turns right)

These generated movement rules represent all the possible subpattern occurences where vehicles have traveled from road segments S_{58} and S_{112} in that same order. As can be observed, the only available next road segments traversed by vehicles in the collected vehicular paths are one of the four mentioned in the movement rules. In order to measure the confidence of a generated movement rule, for example Rule 2, the following equation was used:

$$Confidence_{Rule2} = \frac{Support(S_{58}, S_{112}, S_{19})}{Support(S_{58}, S_{112})} \tag{5.3}$$

Figure 5.8 shows the confidence of each of the possible generated movement rules. The total of the confidences should add up to 100%, as the generated movement rules should be all the possible subpatterns that can occur when a vehicle goes through road segments S_{58} and S_{112} in that same order.

From the calculated confidences for each generated movement rule shown in Figure 5.8, Rule 1 showed a confidence of 7.43%, Rule 2 showed a confidence of 37.16%, Rule 3 showed a confidence of 43.92% and Rule 4 showed a confidence of 11.49%. More specifically, for Rule 2 for example, the subpattern [S_{58}, S_{112}, S_{19}] has occured in 37.16% of all the possible generated movement rules for subpattern [S_{58}, S_{112}]. In order to validate

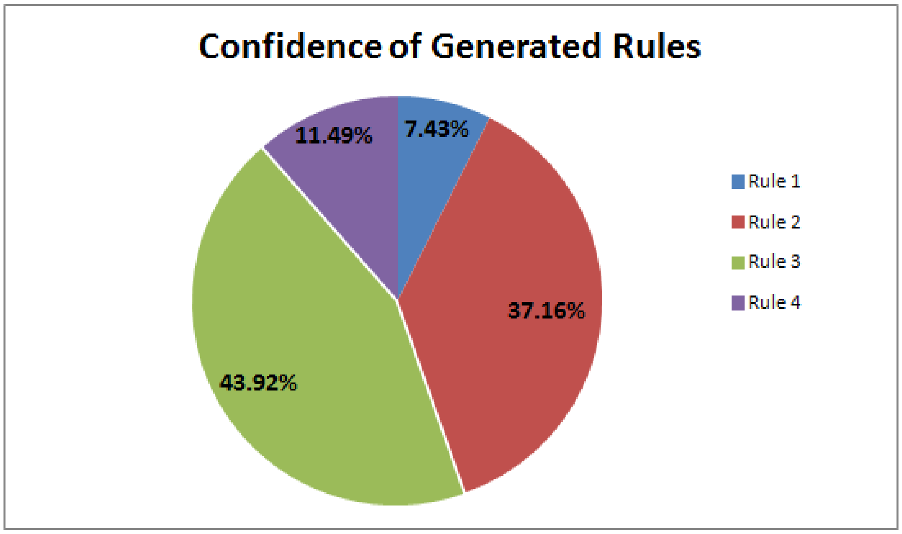

Figure 5.8: Confidence of Generated Movement Rules for a Frequent Movement Pattern

the calculation of the confidences for each rule, the sum of all confidences should equal to 100% as mentioned previously. The sum of confidences calculated for these generated movement rules have added up to 100%.

5.3.3 Analysis Summary

In the analysis of the vehicular paths data mining, two important factors were evaluated; the extraction of frequent vehicular movement patterns based on a user-defined minimum support threshold and the confidence of generated movement rules from the extracted frequent movement patterns.

From the results shown in Figure 5.6, it was apparent that when the minimum support threshold increases, the number of frequent vehicular movement patterns extracted decreases. The value of the minimum support threshold was chosen to vary from 10% to 50% and the number of extracted frequent vehicular movement patterns was recorded. The reason behind choosing this range of values for the minimum support threshold is because if a higher threshold was chosen, fewer frequent patterns will be extracted, which may disregard vehicular paths that have actually been traversed. Even if they are not as frequent as those extracted by a higher minimum support threshold, they are still present and may need to be included to get a better accuracy in the prediction of vehicular behavior.

The analysis of the confidence of generated movement rules for some frequent movement patterns has also been investigated. For each movement rule generated for junction S_{112} in Figure 5.7, a confidence was calculated with respect to all the different subpattern possibilities. If a minimum threshold of 30% was introduced here for example, the only movement rules that will be generated are those for Rule 2 and Rule 3, which leaves out the other two possibilities where vehicles have traversed before.

Chapter 6

Conclusion

Route and mobility prediction has been an open field in the research world for many years. Many studies have looked at different techniques in order to predict the movement of vehicles in a certain geographical area. These techniques have usually required the availability of previous historical trace data which describes the movement of these vehicles. This thesis has presented some important and new definitions of sequential patterns in the context of VANETs, proposed five novel data gathering communication schemes that are responsible for collecting vehicular movement patterns and performed some sequential patterns data mining on the collected paths to extract frequent vehicular movement patterns. The proposed data gathering communication schemes were designed and implemented in the ns-2 environment and their performance and impact on the network was evaluated through a set of extensive experiments. In addition, the data mining aspect of the work has been presented in order to reveal the type of useful information that can be extracted from the collected vehicular paths.

6.1 Conclusion

In more details, this thesis presented a background discussion of the available studies and research in the literature with regards to three topics; vehicular route data gathering techniques, data mining techniques in VANETs, and route and mobility prediction. A summary at the end of each topic was provided to compare the common characteristics they hold. With this information, a broader view of the work done in the literature in relation to the work done in this thesis was achieved.

Next, a set of formal definitions representing vehicular movement behavior was provided. These definitions were customized to fit the context of VANETs as well as represented in a form that can be utilized to extract frequent vehicular movement patterns from them. Important terms such as support and confidence of rules were explained in order to understand their significance during the data mining process.

The thesis then proposed five novel data gathering communication schemes that make use of the infrastructure on the road to collect vehicular paths. The schemes proposed were of two types; RSU-triggered and vehicle-triggered, which depicted the communication trigger for the collection of paths from vehicles on the road. The RSU-triggered scheme periodically queried the vehicles to send their traversed paths to neighboring RSUs, whereas in the vehicle-triggered schemes, vehicles initiated the sending of their traversed paths according to a certain criteria. In addition, the vehicle-triggered schemes were of two modes; broadcast mode and RSU discovery mode. In the former, the vehicles will send their collected paths in a broadcast manner to all RSUs in their vicinity, whereas in the latter, vehicles need to first find the closest RSU available to them before sending their collected paths.

Finally, each of the proposed data gathering communication schemes were evaluated against some important network performance metrics; communication overhead, average packet delay and average packet delivery ratio. The results showed interesting and promising results, which have lead to the understanding of which scheme could be used for which phase of the roadmap shown in 1.2 in Chapter 1. In addition, some data mining was performed on the collected vehicular paths in order to extract frequent movement patterns. The effect of specifying different minimum supports on the extraction process was demonstrated showing that as the support threshold defined was increased, less frequent patterns were extracted, making the data mining process more sensitive to the nature of the frequent patterns. Moreover, the confidence of some movement rules generated for a frequent movement pattern was evaluated showing that, for each rule, a probability of occurence can be extracted. This confidence can then be used in the Prediction and Monitoring phase to forecast the next most probable road segment to be traversed by a vehicle.

6.2 Future Work

In this section, some the future works will be outlined as areas of improvement towards the ultimate endeavor of predicting vehicular mobility and improving the performance of the data gathering communication schemes described in this thesis. A small description of each is provided below:

Prediction Analysis: No prediction of vehicular destination was performed in this thesis, however the data gathering communication schemes and the data mining approach followed could be used to reach this goal. Accuracy and prediction vision (i.e. how many road segments can be forecasted) can both be evaluated and compared to other route and mobility prediction techniques.

Centralized VS Distributed Data Mining: The proposed data mining technique in this work was done in a centralized manner. In other words, all the paths were collected from each RSU into one place and the extraction of the frequent movement patterns was performed. Future work can design and implement a distributed data mining scheme where each RSU can perform data mining operations in order to extract "local" frequent movement patterns in a certain region.

RSU-Awareness: The proposed data gathering communication schemes can be extended in such a way that vehicles can be aware of the RSUs available in a certain region based on their location without the need to discover RSUs. This could significantly increase the performance of those schemes that have been designed to perform under the RSU Discovery mode.

Bibliography

[1] Singh, A.; Kumar, M.; Rishi, R.; Madan, D. K.; "A Relative Study of MANET and VANET: Its Applications, Broadcasting Approaches and Challenging Issues," Communications in Computer and Information Science, 2011, Volume 132, Part 3, 627-632, DOI: 10.1007/978-3-642-17878-8_63

[2] Hartenstein, H.; Laberteaux, K.P.; , "A tutorial survey on vehicular ad hoc networks", Communications Magazine, IEEE , vol.46, no.6, pp.164-171, June 2008 doi: 10.1109/MCOM.2008.4539481

[3] Krumm, J., "A Markov Model for Driver Turn Prediction", SAE Technical Paper 2008-01-0195, 2008, doi:10.4271/2008-01-0195.

[4] Lytrivis, P.; Thomaidis, G.; Amditis, A.; , "Cooperative Path Prediction in Vehicular Environments," Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on , vol., no., pp.803-808, 12-15 Oct. 2008 doi: 10.1109/ITSC.2008.4732629

[5] Ming-Syan Chen; Jong Soo Park; Yu, P.S.; , "Efficient data mining for path traversal patterns," Knowledge and Data Engineering, IEEE Transactions on , vol.10, no.2, pp.209-221, Mar/Apr 1998 doi: 10.1109/69.683753

[6] Simmons, R.; Browning, B.; Yilu Zhang; Sadekar, V.; , "Learning to Predict Driver Route and Destination Intent," Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE , vol., no., pp.127-132, 17-20 Sept. 2006 doi: 10.1109/ITSC.2006.1706730

[7] Andreas D. Lattner, Andrea Miene, Ubbo Visser, and Otthein Herzog. 2006. "Sequential pattern mining for situation and behavior prediction in simulated robotic

soccer." In RoboCup 2005, Ansgar Bredenfeld, Adam Jacoff, Itsuki Noda, and Yasutake Takahashi (Eds.). Springer-Verlag, Berlin, Heidelberg 118-129.

[8] Anwar, F; Petrounias, I.; Morris, T; Kodogiannis, V.; "Discovery of events with negative behavior against given sequential patterns." IEEE Conf. of Intelligent Systems 2010: 373-378

[9] Jesup, E., K. L. Casavant, and C. T. Lawson. "Truck Trip Data Collection Methods." FHWA-OR-RD-04-10. Oregon Department of Transportation, Salem, Feb. 2004.

[10] Azzedine Boukerche, Horacio A. B. F. Oliveira, Eduardo F. Nakamura, and Antonio A. F. Loureiro. 2008. "Vehicular Ad Hoc Networks: A New Challenge for Localization-Based Systems," Comput. Commun. 31, 12 (July 2008), 2838-2849. DOI=10.1016/j.comcom.2007.12.004 http://dx.doi.org/10.1016/j.comcom.2007.12.004

[11] Abumansoor, O.; Boukerche, A.; , "A cooperative multi-hop location verification for Non Line Of Sight (NLOS) in VANET," Wireless Communications and Networking Conference (WCNC), 2011 IEEE , vol., no., pp.773-778, 28-31 March 2011 doi: 10.1109/WCNC.2011.5779230

[12] Tong Liu; Bahl, P.; Chlamtac, I.; , "Mobility modeling, location tracking, and trajectory prediction in wireless ATM networks," Selected Areas in Communications, IEEE Journal on , vol.16, no.6, pp.922-936, Aug 1998 doi: 10.1109/49.709453

[13] Pattara-atikom, W.; Peachavanish, R.; , "Estimating Road Traffic Congestion from Cell Dwell Time using Neural Network," Telecommunications, 2007. ITST '07. 7th International Conference on ITS , vol., no., pp.1-6, 6-8 June 2007 doi: 10.1109/ITST.2007.4295824

[14] Zaidi, Z.R.; Mark, B.L.; , "Real-time mobility tracking algorithms for cellular networks based on Kalman filtering," Mobile Computing, IEEE Transactions on , vol.4, no.2, pp. 195- 208, March-April 2005 doi: 10.1109/TMC.2005.29

[15] Civilis, A.; Jensen, C.S.; Pakalnis, S.; , "Techniques for efficient road-network-based tracking of moving objects," Knowledge and Data Engineering, IEEE Transactions on , vol.17, no.5, pp. 698- 712, May 2005 doi: 10.1109/TKDE.2005.80

[16] Rezgui, J.; Cherkaoui, S.; , "Detecting faulty and malicious vehicles using rule-based communications data mining," Local Computer Networks (LCN), 2011 IEEE 36th Conference on , vol., no., pp.827-834, 4-7 Oct. 2011 doi: 10.1109/LCN.2011.6115558

[17] Fogue, M.; Garrido, P.; Martinez, F.J.; Cano, J.-C.; Calafate, C.T.; Manzoni, P.; Sanchez, M.; , "Prototyping an automatic notification scheme for traffic accidents in vehicular networks," Wireless Days (WD), 2011 IFIP , vol., no., pp.1-5, 10-12 Oct. 2011 doi: 10.1109/WD.2011.6098139

[18] F. Kargl, Z. Ma, E. Schoch, "Security Engineering for VANETs," 4th Workshop on Embedded Security in Cars (escar 2006), November 2006

[19] Dornbush, S.; Joshi, A.; , "StreetSmart Traffic: Discovering and Disseminating Automobile Congestion Using VANETs," Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th , vol., no., pp.11-15, 22-25 April 2007 doi: 10.1109/VETECS.2007.15

[20] Ihn-Han Bae; Olariu, S.; , "A Tolerant Context-Aware Driver Assistance System for VANETs-Based Smart Cars," Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE , vol., no., pp.1-5, 6-10 Dec. 2010 doi: 10.1109/GLOCOM.2010.5684041

[21] Yi Qian; Kejie Lu; Moayeri, N.; , "A Secure VANET MAC Protocol for DSRC Applications," Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE , vol., no., pp.1-5, Nov. 30 2008-Dec. 4 2008 doi: 10.1109/GLOCOM.2008.ECP.376

[22] Fan Li; Yu Wang; , "Routing in vehicular ad hoc networks: A survey," Vehicular Technology Magazine, IEEE , vol.2, no.2, pp.12-22, June 2007 doi: 10.1109/MVT.2007.912927

[23] H. Moustafa, Vehicular networks : techniques, standards, and applications. CRC Press, 2009.

[24] Guangtao Xue; Zhongwei Li; Hongzi Zhu; Yunhuai Liu; , "Traffic-Known Urban Vehicular Route Prediction Based on Partial Mobility Patterns," Parallel and Distributed Systems (ICPADS), 2009 15th International Conference on , vol., no., pp.369-375, 8-11 Dec. 2009 doi: 10.1109/ICPADS.2009.129

[25] Jon Froehlich and John Krumm, "Route Prediction from Trip Observations", Society of Automotive Engineers (SAE) 2008 World Congress, April 2008, Paper 2008-01-0201

[26] Samarah, S.; Al-Hajri, M.; Boukerche, A.; , "A Predictive Energy-Efficient Technique to Support Object-Tracking Sensor Networks," Vehicular Technology, IEEE Transactions on , vol.60, no.2, pp.656-663, Feb. 2011 doi: 10.1109/TVT.2010.2102375

[27] R. Agrawal, R. Srikant, "Mining sequential patterns," Data Engineering, International Conference on, p. 3, 11th International Conference on Data Engineering (ICDE'95), 1995

[28] Rawling, Gerald F. and John P. Reilly. "CATS Commercial Vehicle Survey of 1986: A Discussion of Project Management Issues," CATS Research News. Chicago Area Transportation Study. February 1987.

[29] Ruiter, Earl. "Development of an Urban Truck Travel Model for the Phoenix Metropolitan Area," Final Report. Arizona Department of Transportation. Prepared by Cambridge Systematics, Inc. February 1992.

[30] Barton-Aschman Associates, Inc. "El Paso Urban Area Travel Study, Commercial Truck Travel Survey," Draft Report. Prepared for the City of El Paso Metropolitan Planning Organization and the Texas Department of Transportation. October 1994.

[31] Wilbur Smith Associates. "Commercial Vehicle Survey: Final Report," Prepared for the Houston-Galveston Area Council. 1995.

[32] The Port Authority of New York and New Jersey. "1991 Interstate Truck Commodity Survey," Volume 2. 1992.

[33] Barton Aschman Associates, Inc. "Truck Intercept Survey Procedures Manual," Prepared for Caltrans Alameda County. Barton Aschman Associates, Inc. March 1991.

[34] Gorys, Julius. "1988 Ontario Commercial Vehicle Survey." Transportation Research Record. No. 1313. Transportation Research Board, National Research Council. Washington, DC. 1991.

[35] Sayad, Saed, "An Introduction to Data Mining", Data Mining Group, University of Toronto, 2010, chem-eng.utoronto.ca/ datamining/dmc/data_mining_map.htm

[36] S. Samarah, "Knowledge Discovery for Behavioral Patterns in Wireless Sensor Networks." PhD thesis, University of Ottawa, 2008.

[37] Han, J. and Kamber, M. Data Mining: Concepts and Techniques, 2nd edition. Morgan Kaufmann Publisher, 2006

[38] Kim, M.; Kotz, D.; Kim, S.; , "Extracting a Mobility Model from Real User Traces," INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings , vol., no., pp.1-13, April 2006 doi: 10.1109/INFOCOM.2006.173

[39] Huiping Cao; Mamoulis, N.; Cheung, D.W.; , "Mining frequent spatio-temporal sequential patterns," Data Mining, Fifth IEEE International Conference on , vol., no., pp. 8 pp., 27-30 Nov. 2005 doi: 10.1109/ICDM.2005.95

[40] C. Gao; J. Wang; Y. He; L. Zhou;, "Efficient mining of frequent sequence generators," Proceedings of International Conference on World Wide Web (2008), pp. 10511052

[41] Patwari, N.; Hero, A.O., III; Perkins, M.; Correal, N.S.; O'Dea, R.J.; , "Relative location estimation in wireless sensor networks," Signal Processing, IEEE Transactions on , vol.51, no.8, pp. 2137- 2148, Aug. 2003 doi: 10.1109/TSP.2003.814469

[42] Network Simulator. ns2. http://www.isi.edu/nsnam/ns.

[43] Bisnik, N.; Abouzeid, A.A.; , "Queuing Delay and Achievable Throughput in Random Access Wireless Ad Hoc Networks," Sensor and Ad Hoc Communications and Networks, 2006. SECON '06. 2006 3rd Annual IEEE Communications Society on , vol.3, no., pp.874-880, 28-28 Sept. 2006 doi: 10.1109/SAHCN.2006.288575

[44] Wen Wei; Zhang Heng Yang; , "Increasing Packet Delivery Ratio in GPSR Using Buffer Zone Based Greedy Forwarding Strategy," Data Storage and Data Engineering (DSDE), 2010 International Conference on , vol., no., pp.178-182, 9-10 Feb. 2010 doi: 10.1109/DSDE.2010.17

[45] Yau, K.-L.A.; Komisarczuk, P.; Teal, P.D.; , "A context-aware and Intelligent Dynamic Channel Selection scheme for cognitive radio networks," Cognitive Radio Oriented Wireless Networks and Communications, 2009. CROWNCOM '09. 4th International Conference on , vol., no., pp.1-6, 22-24 June 2009 doi: 10.1109/CROWNCOM.2009.5189427

[46] Kyasanur, P.; Vaidya, N.H.; , "Routing and interface assignment in multi-channel multi-interface wireless networks," Wireless Communications and Networking Conference, 2005 IEEE , vol.4, no., pp. 2051- 2056 Vol. 4, 13-17 March 2005 doi: 10.1109/WCNC.2005.1424834

[47] Jialiang Li; Chunxiao Chigan; , "Delay-Aware Transmission Range Control for VANETs," Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE , vol., no., pp.1-6, 6-10 Dec. 2010 doi: 10.1109/GLOCOM.2010.5684168

[48] Shao Liu, Tamer Baar, R. Srikant, "TCP-Illinois: A loss- and delay-based congestion control algorithm for high-speed networks," Performance Evaluation, Volume 65, Issues 67, June 2008, Pages 417-440, ISSN 0166-5316, 10.1016/j.peva.2007.12.007.

[49] 3rd Generation Partnership Project, "Technical Specification Group Radio Access Network: UTRAN Iub Interface NBAP Signalling," Version: v3.2.0 (2000-06)(Release 1999)

[50] Yazhi Liu, Jianwei Niu, Jian Ma, Lei Shu, Takahiro Hara, Wendong Wang, "The insights of message delivery delay in VANETs with a bidirectional traffic model," Journal of Network and Computer Applications, Available online 30 January 2012, ISSN 1084-8045, 10.1016/j.jnca.2012.01.010

[51] Jay Ayres, Jason Flannick, Johannes Gehrke, and Tomi Yiu. 2002. "Sequential PAttern mining using a bitmap representation." In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02). ACM, New York, NY, USA, 429-435. DOI=10.1145/775047.775109 http://doi.acm.org/10.1145/775047.775109