TRAJECTORY MINING USING UNCERTAIN SENSOR DATA

Seminar Report

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

In

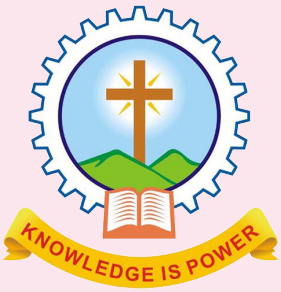
COMPUTER SCIENCE AND ENGINEERING

of

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Submitted By

MISHAL KODALAM POLLATH



Department of Computer Science & Engineering
Mar Athanasius College Of Engineering Kothamangalam

TRAJECTORY MINING USING UNCERTAIN SENSOR DATA

Seminar Report

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

In

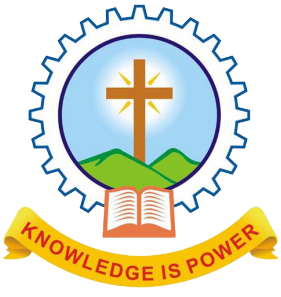
COMPUTER SCIENCE AND ENGINEERING

of

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Submitted By

MISHAL KODALAM POLLATH



Department of Computer Science & Engineering
Mar Athanasius College Of Engineering Kothamangalam

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MAR ATHANASIUS COLLEGE OF ENGINEERING
KOTHAMANGALAM

CERTIFICATE

This is to certify that the report entitled **Trajectory mining using uncertain sensor data** *submitted by* **Mr. MISHAL KODALAM POLLATH**, *Reg.No.***MAC15CS040** *towards partial fulfillment of the requirement for the award of Degree of Bachelor of Technology in Computer science and Engineering from APJ Abdul Kalam Technological University for December 2018 is a bonafide record of the seminar carried out by him under our supervision and guidance.*

.....	.....	.....
Prof. Joby George	**Prof. Neethu Subash**	**Dr. Surekha Mariam Varghese**
Faculty Guide	***Faculty Guide***	***Head Of Department***

Date:

Dept. Seal

ACKNOWLEDGEMENT

First and foremost, I sincerely thank the 'God Almighty' for his grace for the successful and timely completion of the seminar.

I express my sincere gratitude and thanks to Dr. Solly George, Principal and Dr. Surekha Mariam Varghese, Head Of the Department for providing the necessary facilities and their encouragement and support.

I owe special thanks to the staff-in-charge Prof. Joby george, Prof. Neethu Subash and Prof. Joby Anu Mathew for their corrections, suggestions and sincere efforts to co-ordinate the seminar under a tight schedule.

I express my sincere thanks to staff members in the Department of Computer Science and Engineering who have taken sincere efforts in helping me to conduct this seminar.

Finally, I would like to acknowledge the heartfelt efforts, comments, criticisms, co-operation and tremendous support given to me by my dear friends during the preparation of the seminar and also during the presentation without whose support this work would have been all the more difficult to accomplish.

ABSTRACT

Trajectory mining is an interesting data mining problem. Traditionally, it is either assumed that the time-ordered location data recorded as trajectories are either deterministic or that the uncertainty, e.g., due to equipment or technological limitations, is removed by incorporating some pre-processing routines. The trajectories are processed as deterministic paths of mobile object location data. However, the importance is to understand that the transformation from uncertain to deterministic trajectory data may result in the loss of information about the level of confidence in the recorded events. The consideration of uncertain sensor data and transform this to probabilistic trajectory data using pre-processing routines. The data is modelled as tuple level uncertain data and propose dynamic programming-based algorithms to mine interesting trajectories. The results show that the trajectories could be modelled and worked as probabilistic data and that the results could be computed efficiently using dynamic programming.

Contents

List of Figures

List of Tables

List of Abbreviation

IOT	Internet Of Things
GPS	Global Positioning System
RFID	Radio Frequency Identification

Introduction

Trajectory mining is an interesting data mining problem that has been studied in the context of smart cities and the Internet of Things (IoT) [3], [6]. Smart cities and the IoT are indeed the way to the future as trillions of IoT devices, ranging from coffee machines to mobile objects which may or may not be inter-connected, generate enormous amounts of data which need to be modelled and processed effectively to improve daily life [5]. For example, to optimize the commuting time to work, many sources of information including the intended route, calendar, city traffic, weather, etc. need to come together to determine a route which would be the most convenient and therefore, smart data collection, preparation and fast algorithms are needed which can work with the incoming data and propose solutions in real time.

One of the key issues in future smart cities is the incorporation of intelligence into the cities using mobile intelligence [1]. The primary source of mobile intelligence is the mobility data collected through the Internet of Things. The data is obtained from a variety of sources, e.g. moving individuals or devices, which are constantly providing location data, along with a time stamp, to some central repository. Once such data is processed, interesting information could be revealed [2], for instance, which areas in the city are witnessing an increase in activity [3], [4], the location of any traffic anomalies [5], which person or group of people are moving [6], what the popular stay points are [7], etc.

A trajectory is a time-ordered record of a moving object obtained at pre-defined discrete time intervals. However, the 'exact' location of a moving object during these intervals could be uncertain. A lot of research has focused on trajectory uncertainties with an aim to enhance the utility of trajectories. Probabilistic databases offer ways to model uncertainties using possible world semantics [1]. The uncertainties in the trajectories could be at the event level, which is the uncertainty associated with the location of the object, or at the trajectory level, which

is the uncertainty associated with the path recorded as compared to the path taken, or others [8]. An interesting solution in this regard is to record the individual mobile object readings and then create complex events using probabilistic event extraction [5]. Many such systems have been proposed which work with trajectory uncertainty, however Khoussainova et al., proposed creating the events along with confidence values, and this is one of the motivations for the current study.

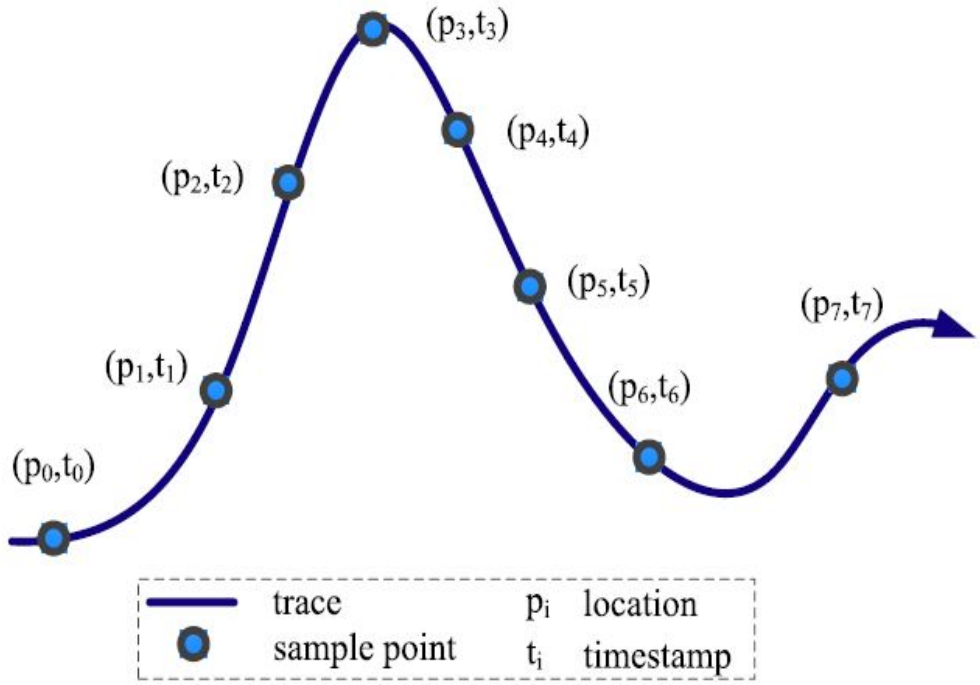


Figure 1.1: A trajectory is generated by sampling from a continuous trace.

A General trajectory figue can be observed from Fig. 1.1. The possible world semantics have been used to model uncertainties. The possible worlds are essentially the all possible combinations of the worlds where an event may or may not be present. However, the idea of using possible world semantics is inherently difficult due to the explosion in the number of possible worlds. Techniques such as dynamic programming have been proposed in the literature which, although giving results which are the same as the possible worlds approach, are significantly

fast. Further, many approximation schemes have been proposed in the literature that give comparable approximate solutions. It should also be noted that in the literature, many simplifying assumptions are made to compute the solution using possible worlds. Such assumptions do not depict real-world situations and thus provide results which may not be very useful. Likewise, relatively complex models tend to be computationally intractable and have been shown to be -P Complete or NP-Complete.

There is a need to develop solutions which model and extract the events in the uncertain trajectory data along with the confidence values and then compute the results based on the confidence values and this is the focus of this work. In this work, we develop a framework that works by collecting the trajectory data obtained from the sensors. The data is stored and processed in a way that helps in identifying events such as key activity areas, evolving activity, etc. thus helping to attain better insight into the work habits of the population. We work on the data obtained from the sensors attached to the office cards. The first step is to pre-process the data and model this as uncertain trajectory data. This is rather challenging considering the types of uncertainties involved, e.g. at the attribute level, i.e. source which generates information, at the tuple level, i.e. the location of the object, or at the trajectory level. Next, the processing that could be performed on the data, which in turn affects the quality of results that could be obtained from the data. Another important aspect is the choice of interestingness measures that is used to compute the results. Many such measures have been proposed in the literature; however, whilst the measures which are simple can be computed in a reasonable time, more detailed measures have been shown to be extremely computing extensive or even computationally intractable. Thus, having a reasonably complex model that can capture the underlying uncertainties in the data as well as a measure which captures the essence of the information in the data is a challenging task which still needs to be investigated at large.

Once the probabilistic models have been developed, a processing framework is required which can be used to answer questions such as, which are the most frequent paths taken by the people in a locality? Or, which are the active regions in a locality over time? etc. The above questions and similar need to be answered in real time and, therefore, efficient mechanisms need to be developed which can work with uncertain trajectory data to yield the live state of

the current happenings in a smart environment. This is extremely challenging as, when the live trajectories are being processed, they should be dealt with as a data stream and while many useful traditional data mining algorithms work under the assumption that the data are stored and could be re-accessed if desired, it becomes extremely challenging to produce similar kinds of results on-the-fly.

Related works

The identification of positioning information regarding people, cars, and other devices is very useful for supporting many of life's daily activities. There are various technologies available to identify location such as Global Positioning Systems (GPS), Radio Frequency Identification (RFID), location estimation of 802.11, GSM beacons, smart phone sensors, infrared or ultrasonic systems. The development of these technologies has made it very easy to produce largescale trajectory data which trace moving objects[3]. Moving objects produce continuous traces in a geographical space from which samples of location points visited by the moving object are taken. A spatial trajectory is an example of trajectory data which include spatial information along with location information. The sampling rate and duration of a trajectory can be chosen depending on the application. In the past decade, many techniques have been developed for trajectory data mining. However, there are several challenges posed by the processing of trajectory data[5]. Typically, the volume of rapidly accumulated trajectory data is large, which makes storage of the data a non-trivial task. Moreover, the definition of a similarity metric for comparing trajectories, which is a basic requirement for trajectory mining, is very important as trajectories may be generated with different sampling strategies considering varying sampling rates[6]. In addition to this, processing queries on large volumes of trajectory data is complicated because of the time and space complexities. Typically, pre-processing is performed on the data which involves cleaning, segmentation, completion, calibration, and sampling. Multiple sensors may detect an object but a deterministic location may not exist. Therefore, cleaning is undertaken which discards the impossible trajectories by considering several constraints such as speed and unreachability constraints. A trajectory may be divided into segments, i.e. sub-trajectories, which correspond to the underlying structures in the data like a path with multiple road segments. The segmentation allows us to efficiently store sample

points of moving objects aligned by time intervals. Low sampling rates may be used for trajectory collection which allow only partial observations of the actual routes because of storage and transmission considerations. Such trajectories are known as uncertain trajectories. Some approaches have already been developed to complete uncertain trajectories and support data mining tasks. Heterogeneous trajectories represent discrete approximations of the original routes and have various strategies and rates of sampling. The heterogeneity may negatively affect the similarity measurement of a trajectory. Therefore, the use of techniques for transformation of heterogeneous trajectories to ones with unified sampling strategies are required.Various sampling techniques are available for reducing a large trajectory database by selecting only those trajectories which accurately represent the original trajectory. A key point is that the sample should capture the hidden mobility pattern from the original trajectory

Proposed work

Trajectory extraction and mining

In this section, we first discuss trajectory generation from the sensor data and then the trajectory mining. A sample trajectory data collection environment is presented in Fig. 1.

3.1 Trajectory data

A trajectory is a collection of location data points ordered by a time stamp. A data point is a triple of the form (eid, sid, e) where eid is the time stamp, sid is the source identifier, and e is the event. The length of a trajectory t is the number of data points it contains,Σ | t|. For example $t =< a_1, a_2, a_3 >$ is a trajectory of three data points ordered in time. A trajectory $s =< a_1, a_2, ...a_n >$ is called a sub-trajectory of a trajectory, $s =< b_1, b_2, ...b_n >$ if $s_i = t_j$ if $s_i = t_j$ for $i, j \in [1, n], i \leq j$ In other words, we say that the trajectory t contains s. Given the location readings obtained from the sensors, a trajectory is obtained by combining in order all the location points recorded for a single object. The trajectory database contains the trajectories for all the sources. The support of a trajectory t is the number of source trajectories that contain the trajectory t. The trajectory mining problem is defined as follows. Given a trajectory database D, find all interesting trajectories, i.e. trajectories whose support is at least a user-specified support threshold θ . Trajectory mining is a multi-stage process which primarily involves pre-processing and pattern mining. We first discuss trajectory data pre-processing.

3.1.1 Data cleaning

Trajectory data is obtained from a variety of sources, e.g. sensors or other mobile devices, and is not entirely correct mainly due to equipment and technological limitations. The errors in trajectory data could be, e.g. (a) a location reading falling out of the motion track (or path) (b) or a moving object recorded at more than one distinct locations, simultaneously. In all such situations, data has to be cleansed. For example, the value of the location attribute is fixed using techniques such as mean/median filters, etc.

3.1.2 Data compression

Trajectory data is recorded at pre-defined discrete time intervals, e.g. a reading every few seconds by each sensor, and most of the points reported are usually repetitive and carry no signfiicant information. However, keeping all the recorded data results in a notable increase in the computational complexity of the problem. It is a common practice to pre-process the data and reduce the number of exact locations recorded, i.e. attain some speed-up at the cost of losing some information which may not be worth the computation effort needed to process the information. Data compression could be offline or on the go and is performed using techniques such as computing the distance metric or similar.

3.1.3 Trajectory creation

The final step is the trajectory creation which typically involves creating a trajectory of time-ordered location data points for each source. Once, the trajectories are created, data mining or other tasks could be performed. Another, important aspect is trajectory data management, however in this work we only focus on trajectory mining. Trajectory mining procedure are shown in Figure 3.1

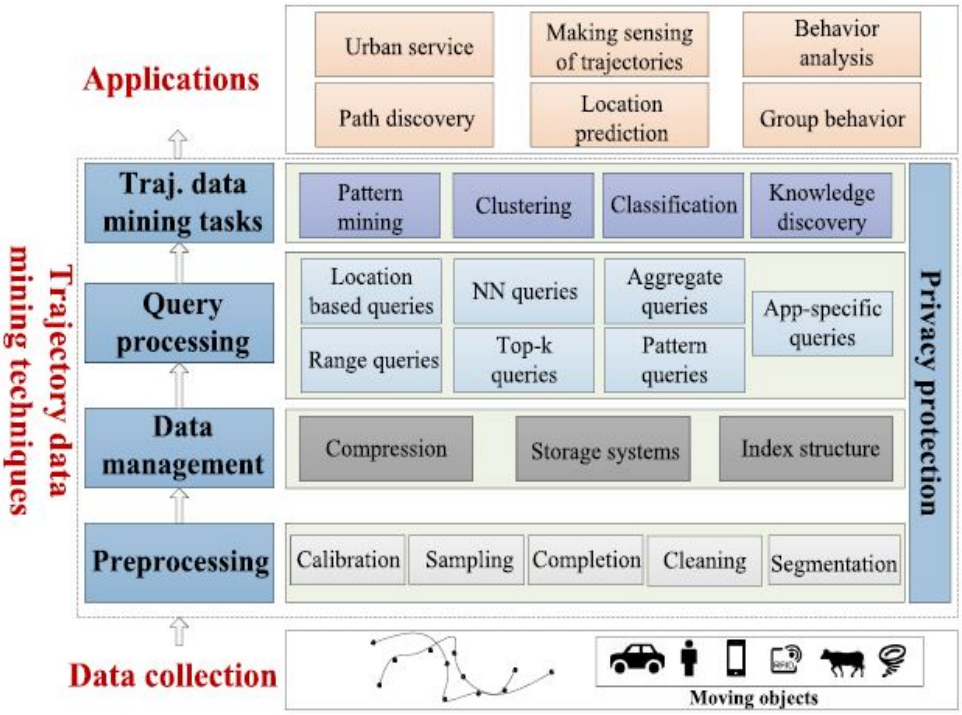


Figure 3.1: A framework of trajectory data mining.

3.2 Trajectory mining

Specifically, the layer of trajectory data mining techniques contains five components listed as follows:

- Preprocessing: In the preprocessing phase, trajectories are usually cleaned, segmented, calibrated, sampled for representatives, or inferred from uncertain trajectories.
- Data management: Sometimes, trajectories are compressed or simplified before being stored. Besides, efficient or scalable storage systems are supposed to be built. Furthermore, appropriate index structures are also necessary to support query processing.
- Query processing: There are various queries that have to be processed to retrieve data, e.g., location-based queries, range queries, nearest neighbor queries, top-k queries, pattern queries, aggregate queries and other application-specific queries. These queries are processed based on an underlying storage system and index structure.
- Trajectory data mining tasks: Trajectory data mining tasks are summarized and classified into several categories, i.e., pattern mining, clustering, classification and knowledge discovery.
- Privacy protection: Privacy-preserving is a crucial problem in every procedure of trajectory data mining. Several examples are provided to illustrate how to process trajectory data as well as to protect sensitive information of users.

Once the trajectories have been created, trajectory patterns are discovered which could be one of the following.

3.2.1 Togetherness patterns

These patterns are aimed at answering questions such as which objects move together. This could help in identifying an emerging activity in a locality or similar. The local administration can use such information in better managing the city's resources, e.g. traffic signals.

3.2.2 Common path patterns

These patterns are the most frequent paths taken by the moving objects. The techniques used for finding such patterns include sequence mining, association mining, etc. These patterns generally help in predicting the next probable location of a moving object.

3.2.3 Group patterns

Similar trajectories are grouped together to find groups of people who move similarly at the same points in time. This is not a trivial task as a feature vector has to be generated which is used to compute the distance between two trajectories. These group patterns show the group mobility trends and could be very useful when dealing with law and order situations, for instance.

3.2.4 Cyclic patterns

Moving objects usually have a similar mobility behaviour over time, i.e. the same activity is performed in cycles, going to work for example. If such personal information is known, it could be used to improve the commuting experience of the individuals, e.g. by warning them of slow moving traffic and then suggesting an alternate route, etc.

3.3 Uncertainty in trajectories

Uncertainty in trajectories is a major concern in many situations as the trajectory data recorded is only a sample of the actual movement. Further, the exact location of a moving object at a specific point in time may not be known. A lot of research has focused on working with trajectory uncertainties. An interesting aspect is to record the object locations along with the confidence values, i.e. along with the location, also record the confidence in an object being at that location. This is a novel idea as in literature the uncertainties associated to the object's location are removed using some threshold based methods and the final trajectory has only deterministic time-ordered data points. This work is focused on working with the uncertainty when dealing with trajectories. In the next section, we discuss the generation of probabilistic

events from trajectory data and then present the probabilistic trajectory mining techniques to extract 'interesting' trajectories from the uncertain trajectory data. The figure 3.2 will explain an uncertain trajectory

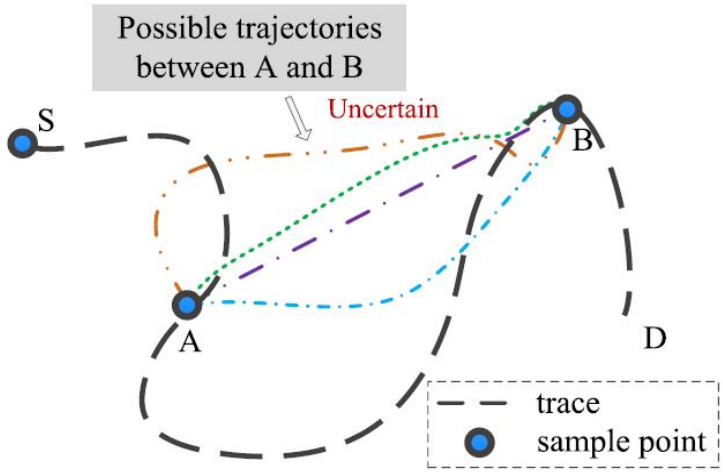


Figure 3.2: An example of uncertain trajectories. A trajectory from S to D is generated at a relatively low sampling rate and only two points A and B are sampled. Movement between A and B is uncertain.

Trajectory mining using uncertain events

Firstly uncertain event extraction from the recorded sensor data should be considered.The data recorded by the sensors is simple location data and the accuracy of such data is usually low. Thus, the events extracted from such data are uncertain, as issue which is discussed below.

4.1 Uncertainity in trajectory path

4.1.1 Uncertain events

The most primitive type of events are presence events.A pres- ence event is of the form $(eid, sid, e, prob)$. For example, a sample reading looks like $(t_1, s_2, l_3, 0.7)$ which means that at time t_1, a source s_2 was spotted at location l_3 with probability 0.7. The reason for having the probability is as follows. For example, a source s1 enters a room which has three (03) antennae installed to detect an object in the room. If two of the three antennae report the presence of the source in the room, there is only a 66.6% chance that the source was in the room at that time. It is also interesting that each antenna records the detection of an object with certainty, i.e. the reading from a single sensor looks like $(t_1, s_2, l_3, 1.0)$ which means that the source s2 was sighted at location l3 at time t1 with probability 1.0. This makes sense as an antenna only reports an event which is detected and there is no uncertainty in this simple event. However, it is the readings from the neighbouring antennae which contribute to the belief in the presence of a source at a specific location. The system takes the readings from multiple antennae and only then decides the confidence in a presence event.

In a subsequent formulation, suppose a source s_1 also enters the room and one out of the three antennae detect s_1. What is the probability of the event that both s_1 and s_2 are in the room, together? An event of this form could be that the sources s_1 and s_2 are at location l_3

with probability 0.18. However, the detection of the sources s_1 and s_2 at location l_3 may not be sufficient to establish that both s_1 and s_2 are at location l_3. There is other information which should also be considered whilst creating such events, e.g. the ownership of the location l_3. If one of the two sources, is the owner of this location, it is probable that the two are together, for example for a meeting. However, if neither of the two owns the location, the detection of these two at the same place with low confidence, i.e. probability 0.18, could be an error. Therefore, the detection and other information are also needed to establish such complex events and for establishing the truth value of a true occurrence.

Further, the sensors are not entirely accurate due to technological limitations, and for example, if a sensor has an error rate of 20% and there are a total of three (03) sensors at a point, then a presence event has a low accuracy.

4.1.2 Uncertain data model

From the previous section, we know that the uncertain events generated by the sensors are of the form (eid, sid, e, prob) which corresponds to tuple-level uncertainty, i.e. a tuple has an existential probability of occurrence. A sample probabilistic trajectory database is shown in Table 4.1. We now define the possible world semantics for an uncertain trajectory database D'.

$time-stamp$	sid	eid	$prob$
1	t_1	u	0.4
2	t_2	w	0.6
3	t_2	v	0.7
4	t_1	v	0.8

Table 4.1: A sample probabilistic trajectory database.

4.1.3 Possible worlds semantics

The possible world semantics are as follows. Given an uncertain trajectory database D', for each event e in a trajectory there are two kinds of worlds: one in which the event is present

and the other where it is not. For each source trajectory t_i, the set of possible worlds is obtained by taking all possible combinations in which an event is present in the world or otherwise. The complete set of possible worlds is obtained by taking all such combinations. The probability of an event occurring is the cumulative probability of occurrence.

$trajectory - id$	$trajectory$
t_1	$(u:0.4)(v:0.8)$
t_2	$(w:0.6)(v:0.7)$

Table 4.2: The trajectory database of table 4.1 transformed to probabilistic trajectories.

$World$	$Probability$
$t_{1,1}$	$<>= (1-0.4) \times (1-0.8) = 0.12$
$t_{1,2}$	$\{u\} = 0.4 \times (1-0.8) = 0.08$
$t_{1,3}$	$\{v\} = (1-0.4) \times 0.8 = 0.48$
$t_{1,4}$	$\{u,v\} = 0.4 \times 0.8 = 0.32$

Table 4.3: The set of possible worlds for source t_1 from table 4.2.

of the worlds where this event is present. As in the literature, we assume that the events across possible worlds occur independently of each other. An example of possible world computation is shown in Tables 4.2-4.4 for the sample database of Table 4.1 transformed to a trajectory database in Table 4.2.

$World$	t_1	t_2	$Pr(D'*)$
D'_1	$<>= 0.12$	$<>= 0.12$	0.12×0.12
$D'_{1,2}$	$<>= 0.08$	$\{v\} = 0.18$	0.08×0.18
....	.....	.....	.....
D'_{16}	$\{u,v\} = 0.32$	$\{w,v\} = 0.42$	0.32×0.42

Table 4.4: Complete set of possible worlds for the trajectory database of table 4.2.

4.1.4 Interestingness measure

Using the possible world semantics, an event that occurs in a signicant number of worlds with high probability is considered an interesting event. The interestingness measure, the expected support of an event, is dened in terms of the expectation of the event occurring in all the possible worlds, i.e. for a trajectory t, the equation 4.1 is explains

$$ExpSup(t, D') = \Sigma_{D^* \in PW(D')} Pr[D^*] * Sup(t, D^*) \qquad ....(Equ : 4.1)$$

For example, the expected support of a trajectory {u,v} in the sample database of Table 4.1 is computed by taking the sum of the probabilities of all the possible worlds which contain {u,v}, i.e. worlds D'$_{12}$ - D'$_{16}$, as shown in Table 4.4.

4.1.5 Uncertain pattern mining

The uncertain pattern mining problem is defined as follows. Given a trajectory database, find all frequent patterns whose expected support is at least a user-defined support threshold θ.

Note that the number of possible worlds is exponential in nature and computing the expected support using possible worlds becomes computationally intractable.We now present a dynamic programming approach to compute the expected support of a trajectory.

4.1.6 Expected support computation

Given a trajectory and a source trajectory, we create a dynamic programming matrix M, $(q + 1) \times (p + 1)$, where q is the number of elements in the trajectory and p is the number of elements in the source trajectory, and initialize all elements in the top row equal to 1 and all elements in the first column (except the top entry) equal to 0. Next, we compute the other values row-by-row by using the following relation. The equation 4.2 explains the dynamic programming matrix:

$$M[i, j] = (1C_{ij}) \times M[i, j - 1] + c_{ij} \times M[i - 1, j - 1] \qquad ....(Equ : 4.2)$$

An example of this computation is shown in Table 4.5. The right bottom cell in the table gives the expected support of the trajectory {u,v}.

		{u:0.4}	{v:0.8}
		1	1
{u}	0	$0.4 \times 1 + (1 - 0.4) \times 1 = 0.4$	0.4
{u,v}	0	0	$0.4 \times 0.8 + (1 - 0.8) \times 0 = 0.32$

Table 4.5: Computing expected support using dynamic programming.

The expected support of a trajectory t in the trajectory database D' is computed by summing the expected support of t across all trajectories.

Algorithm 1 An outline of the trajectory mining algorithm

Given: A Trajectory Database D', An Expected Support threshold Θ
Required: All frequent trajectories

$i = 2$
$F_1 =$: Compute all simple events
while F_{i-1} is not null
C_i =: join F_{i-1} with itself
Prune C_i
for all trajectories in C_i
Compute Expected Support
end for
F_i =: all frequent Ci
i =: i+1
end while
Output the frequent trajectories in F

4.2 Uncertain trajectory mining

The uncertain trajectory mining algorithm is similar to the uncertain apriori algorithm and we give a few details here.An outline of the algorithm is given in Algorithm 1.

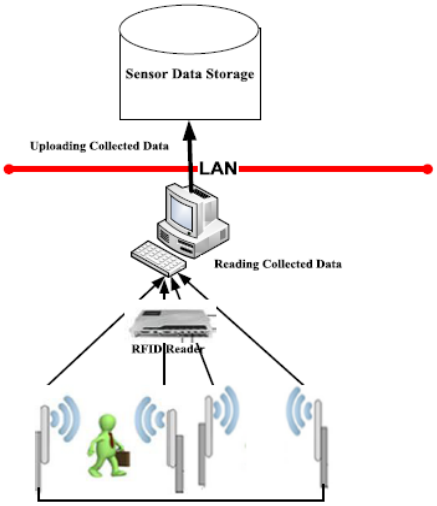


Figure 4.1: A typical sensor-based mobility data collection environment.

4.2.1 Frequent simple events

A scan of the trajectory database D' extracts all simple events in the database D' which have support at least equal to the threshold. The expected support of all the events is computed. Once the database has been scanned, all the frequent simple events have been found and these form the candidates for the next phase, i.e. frequent pair computation. A general schematic representation of trajectory sensing can be observed from Fig. 4.1

4.2.2 Frequent pairs

Once the frequent simple events have been computed, all possible pairs of events are generated which are then tested for being frequent. Note that only the frequent simple events are used to generate candidate frequent pairs. This is due to the apriori property which is anti-monotonic and states that for any pair event to be frequent, both the simple events in the pair have to be frequent. For example, for $\{u, v\}$ to be frequent, both u and v need to be frequent. Only then should the support computation test be performed.

4.2.3 Frequent trajectories

The frequent pairs discovered during the previous phase are used to generate candidate trajectories by appending the frequent simple events to the frequent pairs. The idea is that a candidate trajectory can be extended by appending a simple event to a frequent trajectory which has already been discovered. This step continues until no more candidate trajectories can be created or all the frequent trajectories have been discovered. An outline of the process discussed in Section III is shown in Fig. 4.2. As shown in the figure, it is a two-stage process. In the first stage, the trajectories are extracted which are then mined to extract frequent trajectories in the second stage.

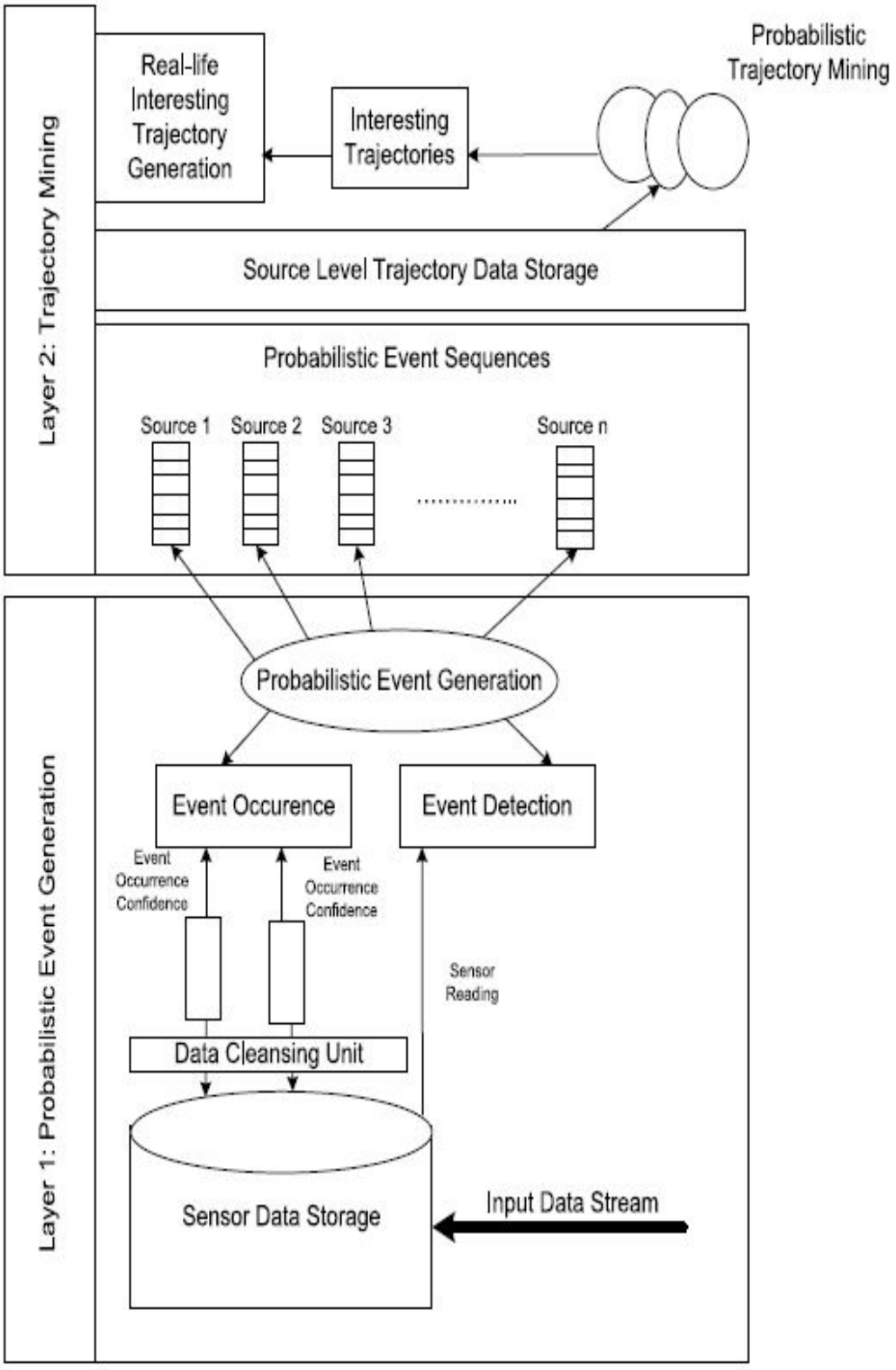


Figure 4.2: An overview of trajectory mining using uncertain sensor data. Layer 1 deals with the probabilistic event generation whilst trajectory mining is performed at layer 2.

Applications of trajectory data mining

A wide spectrum of applications are driven by trajectory data mining. In the section, we classify these applications into following six types. We then introduce each kind of applications through a few examples.

5.1 Applications

5.1.1 Path discovery

Path discovery is one of the most common applications of trajectory data mining. It is extremely important to find the most suitable path in many application scenarios. Exact meaning of the word ”suitable” depends on applications. It can be the fastest, the shortest, the most popular, and so on. Path discovery, also called route discovery, is to find at least one path that satisfy a predefined objective given a source and a destination. Routes must be derived based on a specific road network. Furthermore, geographical locations in numerical style in trajectories should be matched to a map in order to derive candidate paths or path segments. Historical trajectories on the road network provide valuable intelligence to estimate, compare and even construct candidate routes. The fastest path problem is a modification of the shortest path problem. It can be solved by setting edge costs to be time-related factors, e.g., travel time, instead of road distances. However, sometimes the problem is generalized to multiple destinations . The objective is to minimize the cost of a combination of destinations. When planning a trip in an unfamiliar area, people usually try to find the most frequent path between two locations. Furthermore, in a more realistic scenario, an problem is possible to find the most frequent path in a certain time period, i.e., given a time period T , a source vs and a destination v_d , searching the most frequent path during T . Apart from time period constraints, consider

a situation of uncertain trajectories, where trajectories are generated at a very low sampling rate due to multiple reasons, i.e., hardware limitations, privacy concerns, energy constraints. Generally speaking, the most frequent paths outperform the fastest paths or the shortest paths since the frequent ones reflect common routing preferences of previous travelers. It also helps to reduce the risk of failed paths which are possibly unpaved, dangerous or blocked by a recent road work. In terms of public transportation, people's real demand for public transportation are employed to identify and optimize existing flawed bus routes, thus improving utilization efficiency of public transportation. To take into account various driving preferences, a recommendation system chooses different routes for drivers with different driving preferences. This kind of personalized route recommendation avoids flaws of previous unique recommendation and improves quality of user satisfaction.

5.1.2 Location/Destination prediction

Location based services (LBSs), also called location-aware services, are increasingly beneficial to people in urban areas. It has been revealed that human mobility is extraordinary regular and thus predictable. Many location based applications require location prediction or destination prediction to send advertisements to targeted consumers, to recommend tourist spots or restaurants, or to set destinations in navigation systems. Destination prediction is closely related to path discovery. If an ongoing trip matches part of a frequent route in a dataset of historical trajectories, the destination of the frequent route is possibly the destination of the ongoing trip. However, there exist a few constraints in real world scenarios. Research examples are stated as follows. a data sparsity problem, which indicates that available trajectories are too few to cover all possible trajectories. To tackle the data sparsity problem, all trajectories are decomposed into sub-trajectories, and then synthesized trajectories are generated by connecting sub-trajectories together. An expanded set of trajectories that can support destination prediction is exponentially increased by this method. In this paper, privacy protection is also considered to protect sensitive location information of users. Noulas et al. focus on a problem of predicting the next place that a user will visit, by exploring human mobility patterns. A large amount of check-in data are utilized to study human movement with a qualitative representation. Then a

set of features which are corresponding to potential factors that may drive movement of users are extracted. Apart from mobility patterns of individual users, another study further thinks about social conformity of users, i.e., one's movement is influenced by others'. Both regularity and conformity are considered to improve the predictive power. Moreover, heterogeneous mobility datasets e.g., GPS trajectories, cellular tower data, WiFi signals, smart card transactions, check-in locations from online social networks instead of a single type of trajectories are introduced to boost prediction performance.

5.1.3 Movement behaviour analysis

Trajectory data provide a lot of opportunities to analyze movement behavior of moving objects. Discovery of movement patterns is crucial for understanding human behavior. One important challenge in this topic is to extract high-level semantics of behavior, i.e., inferring underlying purposes or roles of moving objects. Renso et al. propose an approach to understand behavior of people who move in a geographical context by extracting mobility behavioral patterns. Then, human behavior is inferred from these patterns which are mined from trajectory data. Predicting human behavior accurately under emergency is a crucial issue for disaster alarming, disaster management, disaster relief and societal reconstruction after disasters. Song et al. analyze emergency behavior of human beings and their mobility patterns after a big nuclear accident in Japan, leveraging a large human mobility database. It is proved that emergency behavior after disasters sometimes correlates with their normal mobility patterns. Furthermore, several impacting factors, e.g., social relationship, intensity of a disaster, damage level, new reporting, population ow, are investigated and thus a predictive model is derived. Another study addresses a problem of detecting roles of moving objects from trajectory data. It is assumed that the intrinsic structure, i.e., the distribution of behavior, characterizes each role. Consequently, the role of a moving object can be identified by exploring structures of trajectories. Human mobility behavior can be studied from spatial, temporal and social aspects. Gao et al. present a comprehensive analysis of temporal effects in modeling mobility behavior. It has been studied that human mobility exhibits strong temporal cyclic patterns in the period of hour, day or week. Liu et al. propose a method to model trajectories in terms of user decision on visiting a point

of interest (POI) and conduct rationality analysis upon trajectory behavior. Rationality of trajectory behavior is explored through several impacting factors. Another recent work explores individual human mobility patterns by studying a large number of anonymous position data from mobile phone users and reveals a high degree of temporal and spatial regularity in human trajectories.

5.1.4 Group behaviour analysis

Moving objects, especially people and animals, sometimes tend to form groups or clusters due to their social behavior. For instance, movement of a person is affected by not only personal activities, but also social ties with that of the groups he belongs to. Besides, a gathering pattern, as a novel modeling of trajectory patterns, describes movement pattern of a group of moving objects. Examples include celebrations, parades, traffic congestion, large-scale business promotions, protests, etc. The topic of mining gathering patterns or group patterns has attracted a lot of research attention. Informally, a gathering in reality indicates an unusual or significant event. a gathering pattern generated by a dense and continuing group of moving objects. Gathering removes requirement for coherent membership in traditional group patterns (e.g., flock, convoy and swarm), leading to a general membership that allows moving objects to enter or leave its group anytime. An extension derives an efficient online discovery approach, i.e., in an incremental manner to incorporating newly generated trajectory data. Another study also aims at efficiently discovering moving objects which move together. A group is defined as a cluster that at least m moving objects being densely connected for at least a certain duration of time. It is very different from gathering meaning aforementioned. Besides, a sampling-independent approach is proposed to avoid flaws of sampling dependent ones, e.g., convoy, swarm. a problem of efficiently modeling individual and group behavior and then present a simulation framework that simulates people's movement behavior in order to generate spatio-temporal movement data. The simulation is of great significance since a large amount of movement data in public domain are limited and unavailable in reality. A recent study is to detect and analyze moving dynamic spatio-temporal regions and their mobility in large sensor datasets. This kind of region often implies locally intense areas of precipitation,

anomalous sea surface temperature readings, and locally high levels of water pollution, etc. It can also be regarded as mining group patterns of a phenomenon.

5.1.5 Urban service

Knowledge discovered with trajectory data mining techniques helps to improve quality of life in urban areas from several aspects Through analyzing a large scale of trajectory data collected from electronic vehicles, charging points, thus minimizing average time to the nearest charging station and average waiting time for an available charging point. Inferring road maps from large-scale GPS traces are highly promising and attractive, since building maps from geographical surveys are expensive and infrequent. address a problem of map inference in a practical setting, i.e., GPS traces has very low resolution and sampling frequency. Several techniques for map inference from sparse data are investigated and extensively evaluated. Traffic volume estimation is a primary task in many applications, such as risk analysis, quality of service, location ranking. A recent study aims to estimate traffic volume for pedestrians within closed environments. Knowledge on people's presence provides a valuable opportunity for improving infrastructure, e.g., locations of information desks, shops or toilets, path-widths of corridors in a stadium. Parking service is of great importance to citizens in urban areas. Parking places (especially on-street parking) are usually unavailable in existing electronic maps. iPark aims to enable parking search applications and to provide complete parking information, i.e., annotating an existing map with parking zones based on trajectory data of vehicles. A developed city naturally has different functional regions, e.g., residential areas, business districts, and educational areas. The knowledge is highly valuable to both citizens and urban planners. People living a city need the knowledge to assist their decision on buying or renting a house, choosing a job. Meanwhile, the knowledge helps urban planners to make decisions on future development of the city and to estimate effects of previous policies. a problem of discovering regions of different functions in a city based on a large scale of trajectory data. A topic model based approach has been proposed to cluster segmented regions into functional zones, where a region is regarded as a document and a function as a topic.

5.1.6 Making sense of trajectories

Raw trajectory data which are in the form of sequence of geographical locations and timestamps fail to make sense to people without semantic description. There exist a great many studies to facilitate interpretation of raw trajectory data. Unlike semantic trajectory that cannot express movement properties of moving objects, e.g., overspeed, stopover,propose a partition-and-summarization approach that automatically generates a short human-readable text to describe a trajectory. The approach not only extends expressivity of traditional semantic trajectories but also avoids a challenging problem of storage, processing and transmission of large volume of semantic trajectories. A raw trajectory data is first segmented according to behavior of a moving object, and then characteristics of each trajectory segmentation are summarized by short textual description. Furthermore, a proto system named STMaker based on this idea is implemented. It is certainly worth noting that semantic meaning of locations and short textual messages collected by social media services provide an unprecedented opportunity to interpret raw trajectory data. detecting a latent topic in trajectory data. Specifically, the approach not only finds semantic regions with a coherent topic but also extracts mobility patterns of human beings between semantic regions. Similarly, a clustering-based approach to discover semantic regions. A lot of emerging location-aware applications require a semantic notation of a location point, e.g., “home”, “work”, instead of latitude and longitude coordinates. Lv et al. propose a method of automatically discovering personal semantic places (i.e., both a physical location and semantic meaning of the location).

Conclusion

In this study, we proposed a framework for probabilistic trajectory extraction and mining from uncertain trajectory data. This is the first study on the subject and many interesting directions need to be explored, e.g. going beyond the number of sources and hours considered in this study. We would also be interested in identifying and developing alternative approaches with the use of which we can make the approach more scalable, e.g. a trajectory compression scheme could be developed to further decrease the length of the trajectories. Further, an approximation scheme could be developed to avoid the dynamic programming processing at the cost of some accuracy. The inherent independence of the trajectories could be used to adapt the proposed algorithm to a distributed computing environment, e.g. Map-Reduce. This work has focused on mining uncertain trajectories. An insight into the discovered trajectories and assessment of the usefulness of the trajectories could also be subjects for interesting future work. We have used expected support as the interestingness measure. Other interestingness measures proposed in the literature, e.g. probabilistic frequentness, could also be investigated and a comparison made with the expected support in terms of time and the discovered trajectories.

REFERENCES

[1] H. J. Levesque, "A logic of implicit and explicit belief," in Proc. AAAI, Aug. 1984, pp. 198202.

[2] N. Khoussainova, M. Balazinska, and D. Suciu, "Probabilistic event extraction from RFID data," in Proc. IEEE 24th Int. Conf. Data Eng. (ICDE), Apr. 2008, pp. 14801482. with stroke for upper limb recovery: A feasibility study," J. Neuroeng. Rehabil., vol. 7, no. 1, pp. 1–17, 2010.

[3] Y. Zheng, "Trajectory data mining: An overview," ACM Trans. Intell. Syst. Technol., vol. 6, no. 3, p. 29, 2015.

[4] P. Wang, Y. Fu, G. Liu, W. Hu, and C. Aggarwal, "Human mobility synchronization and trip purpose detection with mixture of Hawkes processes," in Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2017, pp. 495503.

[5] J. Bao, T. He, S. Ruan, Y. Li, and Y. Zheng, "Planning bike lanes based on sharing-bikes' trajectories," in Proc. 23rd ACMSIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2017, pp. 13771386.

[6] Y. Fu et al., "Sparse real estate ranking with online user reviews and ofine moving behaviors," in Proc. IEEE Int. Conf. Data Mining (ICDM), Dec. 2014, pp. 120129.

[7] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2012, pp. 186194.

[8] D. Suciu, D. Olteanu, C. Ré, and C. Koch, "Probabilistic databases," Synthesis Lectures Data Manage., vol. 3, no. 2, pp. 1180, 2011.