Frequent Route Based Continuous Moving Object Location- and Density Prediction on Road Networks

Győző Gidófalvi*
KTH Royal Inst. of Technology
gyozo.gidofalvi@abe.kth.se

Manohar Kaul
Uppsala University
manukaul@acm.org

Christian Borgelt
EU Centre for Soft Computing
christian@borgelt.net

Torben Bach Pedersen
Aalborg University
tbp@cs.aau.dk

ABSTRACT

Emerging trends in urban mobility have accelerated the need for effective traffic prediction and management systems. The present paper proposes a novel approach to using continuously streaming moving object trajectories for traffic prediction and management. The approach continuously performs three functions for streams of moving object positions in road networks: 1) management of current evolving trajectories, 2) incremental mining of closed frequent routes, and 3) prediction of near-future locations and densities based on 1) and 2). The approach is empirically evaluated on a large real-world data set of moving object trajectories, originating from a fleet of taxis, illustrating that detailed closed frequent routes can be efficiently discovered and used for prediction.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: [Data mining, Spatial Databases and GIS]

General Terms

Algorithms

Keywords

spatio-temporal data mining, mobility patterns, frequent routes, traffic prediction

1. INTRODUCTION

The rapid growth of demand for transportation, and high levels of car dependency caused by the urban sprawl, have exceeded the slow increments in transportation infrastructure supply in many areas. Recently, the wide-spread adoption of GPS-based on-board navigation systems and location-aware mobile devices have enabled a gamut of traffic prediction and management systems to efficiently utilize existing infrastructure and combat the ever increasing gap between the rapid growth of vehicles and the slow increments in infrastructure development.

Consider the following scenario: There are vehicles traveling between *Uppsala* and *Stockholm* (cities in Sweden), each equipped with on-board GPS systems and *map-matching* clients that can align the GPS location coordinates to an appropriate position on a road segment in the road network. Tuples consisting of the vehicle ID and the road segment ID are constantly *streamed* to a central server, giving the server the ability to track vehicle movements via *trajectories* that evolve with time. Assuming these capabilities, is it then possible to predict where a given vehicle will be in exactly 10 minutes from now? What is the estimated number of vehicles on each road segment in the network i.e. road network density, in 10 minutes from now? In case of an accident at a junction in *Knivsta* (intermediate town), can we compute the exact set of vehicles that will arrive here in the next 10 minutes and notify only this set of vehicles, so they can make alternate routing decisions to their respective destinations? Answering the aforementioned questions with acceptable accuracy forms a solid foundation for more effective urban traffic planning techniques.

The contributions of the paper are as follows. First, the paper defines a formal framework for modeling, online mining and prediction of road network based continuously evolving trajectories of moving objects (Sec. 3). Second, a novel prediction model is proposed to predict the near-future location of objects based on historical closed frequent routes. Third, using the proposed model, *continuous queries* (CQs) are formulated in a DSMS to calculate probabilistic future locations and the resulting network density of moving objects [13] (Sec. 4). Finally, the paper empirically evaluates and shows the effectiveness and feasibility of the approach on a large real-world data set of moving objects (Sec. 5).

2. RELATED WORK

A lot of previous work has considered mining of frequent routes from moving object trajectories. This line of work can be broadly classified as *road network based* [1, 4] and *Euclidean space based* [3, 9]. Overall, none of these approaches consider the online mining of closed frequent routes in networks, which is a distinguishing feature of the present paper. Much work has also been devoted to discovering

*This work was partially supported by TRENoP, the strategic research program in Transport at KTH.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ACM SIGSPATIAL GIS '11, November 1-4, 2011. Chicago, IL, USA
Copyright 2011 ACM ISBN 978-1-4503-1031-4/11/11 ...$10.00.

various other types of movement patterns besides frequent routes, both for static and streaming data [2, 6]. The application of mined movement patterns for traffic management has also received much attention recently [5]. Several papers suggest using historical frequent route information to predict near-future locations. Work in [12] assumes that final destinations of all vehicles are known apriori. Approaches in [10] cluster trajectories in a moving object database, form higher granularity *dense regions*, and predict movement between these regions. In the area of network density prediction [7, 13], statistical approaches based on *short-term* observations of traffic movement have mostly been employed. The closest to the present proposal is the work in [11] in which the proposed *network mobility model* consists of (i) turning patterns/statistics at road junctions and (ii) mined travel speeds on road segments. In comparison, the present proposal uses both historical closed frequent routes and turn statistics with speed profiles.

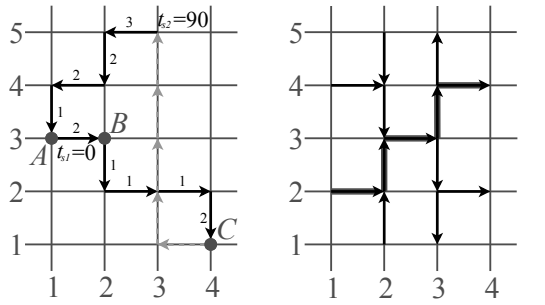
3. PRELIMINARIES AND DEFINITIONS

Road Network Based Routes of Moving Objects: Let $O = \{o_1, \ldots, o_M\}$ be a set of moving objects. Let the time domain be denoted by $\mathbb{T}$ and be modeled as the totally ordered set of natural numbers $\mathbb{N}_0$. Following [1], let the road network be modeled in terms of a set of *base points*, $B \subset \mathbb{R}^2$, a set of *line segments*, $LS \subset B^2$, and *connections*, $C \subset B$. Then, the continuous movement of an object on the road network is described with a *(road network based) trajectory*, $tr^o = (ts, s)$, where ts denotes the starting time of the trajectory and $s = \langle (ls_1, \Delta t_1), \ldots, (ls_m, \Delta t_m) \rangle$ is a *temporally annotated sequence*, i.e., a sequence of pairs of traversed segments, $ls_i \in LS$, and associated *traversal times*, Δt_i.

Trip Trajectories of Moving Objects: Extended stops in movement naturally subdivide the trajectory tr^o of an object $o \in O$ into a sequence of *trip trajectories* $\langle tr^o[1], .., tr^o[t] \rangle$. A trip trajectory $tr^o[i]$ is modeled in the same way as an object trajectory. Figure 1(a) shows an object's trip trajectories on a street grid. A connection is referenced by the concatenation of its coordinates, and a directed segment is referenced by the concatenation of the references of its starting and the ending base points, e.g., the directed segment from A to B is referenced as 1323. The paths along solid black arrows show 2 trip trajectories, $tr^o[1]$ and $tr^o[2]$, of an object's; path along dashed gray arrows represent unobserved trip trajectories. Using the referencing system, the first trip trajectory, $tr^o[1]$, is represented by the pair $(0, \langle (1323, 2), (2322, 1), (2232, 1), (3242, 1), (4241, 2) \rangle)$.

Continuously Evolving Trajectories: As an object $o \in O$ moves, its trip trajectory $tr^o[t]$ is *evolving*, i.e., it is continuously extended at the end of the sequence by appending the segment that o has most recently traversed. A single extension of $tr^o[t]$ is referred to as a *trajectory piece* and the i-th trajectory piece is denoted and modeled as $tp_i^o[t] = (ts_i, (ls_i, \Delta t_i))$. $tp_i^o[t]$ is implicitly associated with an *arrival time*, $t\_arr = ts_i + \Delta t_i$. A sequence of trajectory pieces $\langle tp_i^o[t], \ldots tp_k^o[t] \rangle$ of a trip trajectory $tr^o[t]$ of object o for trip t forms a *contiguous trip sub-trajectory* of object o for trip t if $\forall j$ such that $i \leq j < k$, $ts_j + \Delta t_j = ts_{j+1}$.

Route Mining: The frequent route mining task can be formulated similar to that of the simplified sequential pattern mining task as follows. Let $O = \{o_1, \ldots, o_M\}$ denote a set of objects and let $TR = \{tr_1, \ldots, tr_T\}$ denote the set of their trip trajectories in which tr_i represents a particular trip trajectory $tr^{o_j}[t]$ of object $o_j \in O$ for trip t. A route r (a temporally annotated sequence) is a *Contiguous Frequent Route* (CFR) iff r is *contiguously supported* by at least *min_sup* trajectories. A trajectory $tr_i = (ts_i, s_i) \in TR$ *contiguously supports* r, iff there exist a *contiguous* index sequence $1 \leq i_1 < \ldots < i_l \leq m$ such that $i_{j+1} - i_j = 1 \ \forall j$ where $1 \leq j < l$ and $ls'_j = ls_{i_j} \ \forall j$ where $1 \leq j \leq l$. A route r_c is a *Closed Contiguous Frequent Route* (CCFR) iff r_c is a CFR and there exists no *extended* CFR, r_e, such that r_c is a proper subsequence of r_e, and the support of r_c is equal to the support of r_e. The temporal annotation of the route is defined to be a sequence aggregate (e.g., arithmetic average) of the traversal times of the corresponding segment of the trajectories that support the route. Then the *Closed Contiguous Frequent Route Mining* problem is defined as: Given a set of objects O, a set of their trip trajectories TR, and a minimum support threshold value *min_sup*, find the set of CCFRs in TR. The application-relevant route contiguity and closedness constraints ensure that CCFRs (i) are an optimal, lossless compression of all routes and (ii) contain no gaps and thus can be effectively used for prediction.



(a) Trip trajectories of a moving object.

(b) Sample trajectories of set of moving objects.

Figure 1: Moving object trajectories.

Route Based Prediction: Using the extracted CCFRs, the *Moving Object Location Prediction* task is defined as: Given a road network (B, LS, C), a *set* of CCFRs, and the contiguous trip sub-trajectory $\langle tp_i^o[t], \ldots tp_k^o[t] \rangle$ of object $o \in O$ for its current trip t up to the current time t_c, for all segments, $ls_i \in LS$, calculate the probability, $\Pr(ls_i^o | t_p)$, that o will be located on ls_i at the *prediction time* $t_p \geq t_c$.

Aggregating the location predictions, the *Road Network Density Prediction* task is defined as: Given a road network (B, LS, C), a *set* of CCFRs, and the contiguous trip sub-trajectories of a *set* of object O up to the current time t_c, for all segments $ls_i \in LS$ calculate the expected number of objects, $\mathrm{E}(ls_i | t_p) = \sum_{o \in O} \Pr(ls_i^o | t_p)$, that will be located on ls_i at the prediction time $t_p \geq t_c$.

The set-based tasks of *CCFR* mining, moving object location prediction, and road network density prediction are naturally extended to a *continuous stream of timestamped trip trajectory pieces of objects* by adopting the commonly used temporal sliding window model $SW = (t_{wsize}, t_{wstride}, t_{wlag})$. Detailed definitions are omitted due to space limitations.

4. CCFR-BASED PREDICTION MODEL

The following section describes the proposed prediction model based on past CCFRs with a running example.

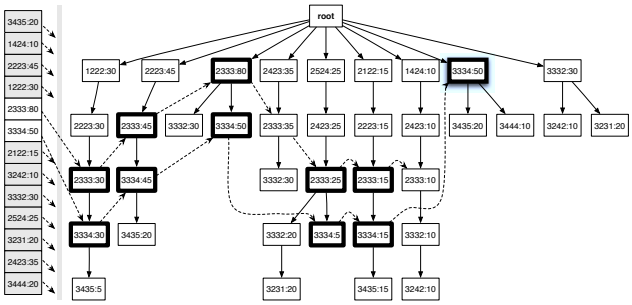


Figure 2: Complete prefix tree with highlighted anchor and other overlapping segments. For qv = {2333,3334}, there is a total overlap in the highlighted branches except the last highlighted branch which has only the anc = 3334.

Running Example Listing 1 shows 7 sample trajectories, which are traced by a number of objects (shown in parenthesis) that are moving on the sample grid road network introduced in Section 3. In addition to the referencing convention of directed segments introduced in Section 3 (Figure 1), at the end of each trip trajectory a segment ID of -1 is used to denote a special virtual segment that is used to signal the end of the trip.

```
T1  1222 2223 2333 3334 3444 −1   (20)
T2  1222 2223 2333 3334 −1        (5)
T3  1222 2223 2333 3334 3435 −1   (5)
T4  2122 2223 2333 3334 3435 −1   (15)
T5  1424 2423 2333 3332 3242 −1   (10)
T6  2524 2423 2333 3334 −1        (5)
T7  2524 2423 2333 3332 3231 −1   (20)
```

Listing 1: Sample trajectories on road network.

CCFR Mining CCFR mining works by growing CCFRs (or patterns) in a depth-first fashion. The direct check of pattern extensions is adopted in the present CCFR mining method. Mining the sample trajectories with *min_sup* set to 10% generates (i) *CCFR* and (ii) all possible turning patterns at road junctions, termed as *turn statistics*, as is shown in Listing 2.

```
P1  {2524,2423,2333,3334}          (5)
P2  {2524,2423,2333,3332,3231}     (20)
P3  ...
```

Listing 2: Sample CCFRs.

Closed Frequent Pattern Tree Creation The mined CCFRs are stored in a prefix tree similar to an *FP-tree* [8] consisting of a prefix tree τ_c and a header table H_c (which acts as an index into the prefix tree). Patterns are read in one at a time and inserted into a prefix tree to enable quick retrieval of CCFRs/patterns similar to [14]. Figure 2 shows the complete prefix tree τ_c (without all *turn statistics*) after inserting all example CCFRs.

CCFR-Based Prediction The proposed prediction model is based on the notion of a *query vector*, qv, a sequence of road segments that a vehicle has traversed, e.g. {2333,3334}, and the *anchor*, anc, the last, most recently traversed segment in qv, i.e. 3334 in the example. Then, the prediction algorithm is composed of four core tasks:

Table 1: Probabilities calculated from CCFRs.

Rule	Calculation	Prob	Left
{2333, 3334} -> 3435	1.0 * (20/50)	0.4	0.6

Table 2: Probabilities from turn statistics.

Rule	Calculation	Prob	Left
{2333, 3334} -> 3444	0.6 * (10/50)	0.12	0.48

***Task 1*:** Finding the probability distribution for the possible next segments given a query vector (qv). For this a "*best match*" to the query is sought. The "best" choice is computed using a cost model that either favors partial matches close to or far away from anc in qv. If there is no CCFR that matches qv at least partially, matching falls back on *turn statistics*. Note here that by storing all occurring segment pairs, the turn statistics are always found to be the best match (thus simplifying the prediction procedure). Highlighted segments in Figure 2 show total matches with $qv = \langle 2333, 3334 \rangle$, except a partial match in the last highlighted branch.

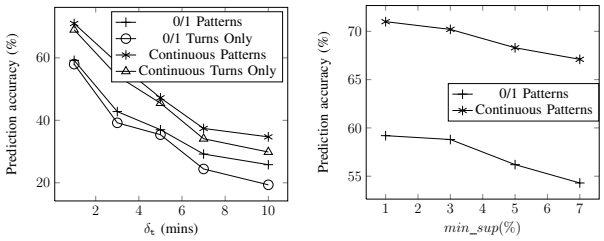
***Task 2*:** In order to estimate probabilities for Task 1, support values are calculated, which is a non-trivial task when dealing with only *closed patterns*. While storing all *frequent patterns* would considerably simplify the task of support calculation, it would require a lot more storage space. From the set of "best" match branches, all child nodes below anc are considered. Rules of the form $sseq(qv) \rightarrow s$, where s denotes child node extensions and $sseq$ denotes a subsequence, for different values of s and a best match of a subsequence of qv, are formed. Probability is then computed as: $P(s|sseq(qv)) = sup(sseq(qv)+s)/sup(sseq(qv))$. Table 1 shows the probabilities and the remaining probability mass that is calculated for each unique child node below anc, i.e., 3435. The remaining probability mass of 0.6 is then distributed using *turn statistics*, shown in Table 2.

***Task 3*:** The query vector qv is extended by a predicted segment and weighted with the probability of the segment, then the prediction is repeated recursively (for each possible next segment). At the time horizon, the probability mass is assigned to the anchor of the extended query. This yields a probability distribution over the segments, describing where the object could be at the time horizon. In the example, an initial call to the prediction algorithm is: $predict(\langle 2333, 3334 \rangle, 2.0, 1.0)$, which implies that a moving object with initial probability mass of 1.0 has traced a path containing 2333 and 3334. At current time t_c, it is located at the end of the anchor segment 3334 and a location prediction is required on time $t_c + 2.0$ time units. The location probabilities for the possible next segments 3435, 3444, and −3334 at recursion depth 1 are calculated to be 0.40, 0.12, 0.48, respectively.

***Task 4*:** The probability distribution over segments at the time horizon is aggregated over all objects in order to obtain a density estimate of both *moving* and *parked* moving objects on the road network.

5. EXPERIMENTS

The proposed method is evaluated on a one day long sample of the near real-time stream of raw GPS positions of 1500 taxis and 400 trucks moving on the streets of Stockholm [15].



(a) Prediction accuracy for varying δ_t values.

(b) Prediction accuracy for varying *min_sup* values.

Figure 3: 0/1 discrete and continuous prediction accuracy for individual moving objects averaged over windows for fixed values of t_{m_wsize} = 3600 seconds (1 hour), $t_{m\_wstride}$ = 1800 seconds (30 minutes), and *min_sup* = 1%.

In the sample, vehicle positions are read approximately once every minute. The status information and spatio-temporal analysis of individual trajectories is used to identify approximately 17,000 trips. To adapt the raw GPS data set to the proposed framework, road network based trajectories are constructed using a subset of the Stockholm road network (approx. 6,000 directional segments with an average segment length of 55 meters and degree of connectivity of 2.3). Three groups of experiments evaluated (i) the throughput and execution time of CCFR steam mining, (ii) the scalability of the CCFR mining method, and (iii) the CCFR-based prediction accuracy.

In the first set of experiments, for low support thresholds (1-2%), fixed mining window stride ($t_{wstride}$) and increasing mining window size, a large increase in the number of mined CCFRs was noted while execution time was still in real-time processing limits, i.e. runtime $< t_{wstride}$.

The second set of experiments measure the scalability of the CCFR mining method w.r.t the input size. Increasingly large volumes of simultaneously moving object trajectories are simulated by fixing the mining window stride $t_{wstride}$ to a 24hr period and increasing the mining window size t_{wsize}. The execution time for nearly 17K input trajectories mined at a minimum support of 0.1% is approximately 40 seconds with approximately 25K CCFRs mined.

In order to measure the accuracy of the CCFR-based prediction approach experiments were carried out to measure the effect of gradually incrementing either the time horizon for location prediction, i.e., $\delta_t = t_p - t_c$, while keeping the minimum support threshold *min_sup* fixed and vice versa. Discrete (0/1) accuracy implies an exact match between predicted and actual road segment while continuous accuracy reflects a value for partial probabilities too. Figure 3 shows a gradual decline in accuracy as δ_t and *min_sup* are increased one at a time. The proposed approach outperforms the "turn statistics only" approach and its additional utility becomes increasingly pronounced as he time horizon is increased.

6. CONCLUSIONS AND FUTURE WORK

The present paper proposed a novel approach to using continuously streaming moving object trajectories for traffic prediction and management. Founded on realistic real-world application requirements, the paper proposes concrete methods, data structures, and a prototype implementation in a DSMS for managing, mining, and predicting the incrementally evolving trajectories of moving objects in road networks. The approach is empirically evaluated on a large real-world data set of moving object trajectories, originating from a fleet of taxis, illustrating that detailed CCFRs can be efficiently discovered and used for prediction.

Future work is planned along several directions. Firstly, the per-object based parallelization and distributed processing of online CCFR-based prediction will be explored. Second, the aggregated matching of groups of similar trajectories against groups of similar CCFRs will be explored. Third, the use and benefits of more sophisticated cost models, e.g., based on the discriminative power of segments or CCFRs, in the CCFR based prediction model will be explored. Finally, the aggregation of CCFRs from different historical mining windows will be explored.

References

[1] A. Brilingaitė and C. Jensen. Enabling routes of road network constrained movements as mobile service context. *Proc. of GeoInformatica*, pages 55–102, 2007.

[2] S. Dodge, R. Weibel, and A.-K. Lautenschütz. Towards a taxonomy of movement patterns. *Information Visualization*, pages 240–252, 2008.

[3] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. *Proc. of SIGKDD*, pages 330–339, 2007.

[4] G. Gidofalvi and T. Pedersen. Mining long, sharable patterns in trajectories of moving objects. *Proc. of GeoInformatica*, 13:27–55, 2009.

[5] G. Gidofalvi and E. Saqib. From trajectories of moving objects to route-based traffic prediction and management. *Proc. of GIScience*, 2010.

[6] G. Gidofalvi, T. Pedersen, T. Risch, and E. Zeitler. Highly scalable trip grouping for large scale collective transportation systems. *Proc. of EDBT*, pages 678–689, 2008.

[7] M. Hadjieleftheriou, G. Kollios, D. Gunopulos, and V. Tsotras. Online discovery of dense areas in spatio-temporal databases. *Proc. of SSDT*, 2003.

[8] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. *Proc. of SIGMOD*, pages 1–12, 2000.

[9] J. Han, J. Lee, H. Gonzalez, and X. Li. Mining massive rfid, trajectory, and traffic data sets. *Proc. of SIGKDD*, 2008.

[10] H. Jeung, Q. Liu, H. Shen, and X. Zhou. A hybrid prediction model for moving objects. *Proc. of ICDE*, pages 70–79, 2008.

[11] H. Jeung, M. Yiu, X. Zhou, and C. Jensen. Path prediction and predictive range querying in road network databases. *Proc. of VLDB*, pages 585–602, 2010.

[12] S. Kim, J. Won, J. Kim, M. Shin, J. Lee, and H. Kim. Path prediction of moving objects on road networks through analyzing past trajectories. *Proc. of KES*, pages 379–389, 2007.

[13] H. Kriegel, M. Renz, M. Schubert, and Z. A. Statistical density prediction in traffic networks. *Proc. of SDM*, 2008.

[14] F. Pinelli, A. Monreale, R. Trasarti, and F. Giannotti. Location prediction within the mobility data analysis environment deaedalus. *Proc. of MobiQuitous*, 2008.

[15] M. Rahmani, H. Koutsopoulos, and A. Ranganathan. Requirements and potential of gps-based floating car data for traffic management: Stockholm case study. *Proc. of ITSC*, pages 730–735, 2010.