Interactive Visual Exploration of Most Likely Movements*

Can Yang and Győző Gidófalvi

1 Introduction

The widespread adoption of location-enabled devices and the increasing acceptance of services that leverage (personal) data as payment enable the collection of massive streams of trajectories of moving objects. Trajectories provide explicit evidence about the route choices of travelers, which is a critical aspect of transport models in transport planning. Yet most transport models, to escape the daunting task of analyzing time-varying route choices for all origin-destination pairs, unrealistically and despite the evidence make the shortest- or least cost path assumption between origins and destinations. To remedy this situation, the paper proposes a method and system that adopts the general visual analytics paradigm of "extraction, depiction, and visualization of computationally extracted patterns" (Andrienko et al., 2008, 2010). In particular, the proposed method and system 1) in an incremental fashion aggregates the time-varying movement information as closed contiguous frequent routes from a massive trajectory stream and 2) reconstructs from this information the *k* most likely movements for a selected origin-destination pair and time period. A simple 2D map interface is used to define the spatial (and temporal) predicates of the user-query and to display and explore the reconstructed movements. The virtues of the proposed system are that 1) it compresses the infinite stream of trajectories to a finite storage space (thereby allowing the efficient processing of queries) and 2) in comparison with a pure data warehouse solution (Krogh et al., 2013, 2014), it can construct movements that might not have been observed in the sample but are likely

Can Yang
KTH Royal Institute of Technology, DKV 30, 10044 Stockholm, Sweden, e-mail: cyang@kth.se

Győző Gidófalvi
KTH Royal Institute of Technology, DKV 30, 10044 Stockholm, Sweden, e-mail: gyozo@kth.se

* This document has been prepared in LaTeX according to the template of AGILE 2016 as suggested by the program committee of VCMA 2016.

in the population. The system is demonstrated using 2.26 million trip trajectories of 11 thousand taxis over a six day period in Wuhan, China.

2 Problem Definition

Let $O = \{o_1, \ldots, o_N\}$ be a set of moving objects. Let the time domain be denoted by T. Let S be a discretization of the 2D space into application specific spatial elements $s_1, s_2, \ldots$, e.g., road network segments, grid cells, regions, where the objects can be located. Without loss of generality assume that S is a unit grid. Let the continuous movement of an object o during its trip[2] on S be modeled as a pair $tr = (t_0, p)$, where t_0 is the start time of the movement and $p = \langle (s_1, \Delta t_1), \ldots, (s_n, \Delta t_n) \rangle$ is a *spatially contiguous sequence of element traversals* $(s_i, \Delta t_i)$, where Δt_i denotes the time it took o to traverse s_i during its trip. Trips of objects are observed as a time-ordered *stream of trip elements* of objects $STE = \langle te_1, te_2, \ldots \rangle$, where a te_i is a three-tuple $(ts_i, (s_i, \Delta t_i), o_i)$ and records the fact that at time ts_i object o_i traversed the spatial element s_i in Δt_i time units.

Let $dow : T \mapsto \{1, \ldots, 7\}$ denote the day-of-week- and $tod : T \mapsto \{0, \ldots, 23\}$ denote the time-of-day temporal domain projection functions. Let P_T be a temporal predicate over T that is defined in terms of temporal domain projected values as $P_T \subseteq \{1, \ldots, 7\} \times \{0, \ldots, 23\}$. Let P_S be a spatial predicate that is defined in terms of an origin region $r_o \subset S$ and a destination region $r_d \subset S$. Then, given the movements of a sample of objects in a population in *STE* and the spatial and temporal predicates P_S and P_T, the *k-Most Likely Movements* (k-MLM) problem is defined as estimating the k most likely movements / paths of the population from the origin region $P_S.r_o$ to the destination region $P_S.r_d$ during the periods defined by P_T.

3 Method

Since the k-MLM problem is defined for the movements of a population of objects based on the movements of a sample, which might not even contain movements that satisfy the spatial and temporal predicates, the proposed approach first incrementally aggregates general movement information of the sample objects for different temporal domain projections as Closed Contiguous Frequent Routes (CCFRs) (Bachmann et al., 2013), then constructs a generative probabilistic movement model based on topological relationships between CCFRs, and expresses the k-MLM problem as extracting k least cost distinct paths problem in the directed transitions graph of CCFRs. Figure 1 shows the stages of the method and the following subsections detail the stages.

[2] A trip is defined as a purposeful movement of an object from an origin to a destination where the object is stationary for an extended period of time.

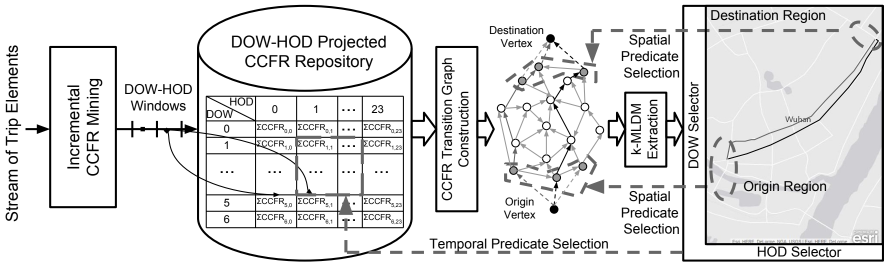


Fig. 1 Schematic diagram of the *k*-MLDM method.

3.1 Incremental Extraction of CCFRs

The proposed method extracts and compresses information about the time-varying movements of a collection of objects in the form of temporally domain projected CCFRs as it is defined in (Bachmann et al., 2013). Informally, given a set of trips *TR* the *support* of a spatially contiguous sequence of element traversals *r*, denoted as $sup(r)$, is the number of trips in *TR* that contain *r*. *r* is *frequent* if $sup(r) \geq min\_sup$. *r* is *closed* if there does not exists another sequence r^* that contains *r* and has the same support as *r*. *r* is a CCFR if it is frequent and closed. To extract time-varying CCFRs from *STE* the proposed method forms 1-hour long tumbling windows over *STE*, extracts CCFRs from the windows, and incrementally aggregates the CCFRs for each *dow* and *hod* combination using the approach of (Bachmann et al., 2013).

3.2 CCFR Based Movement Model

The proposed model assumes that objects probabilistically move along CCFRs and the movement is complete, i.e., starts from the first- and ends at the last spatial element of the CCFR. At the end of a CCFR an object either probabilistically transitions to “connected” CCFRs or stops moving.

Initial probabilities of and transition probabilities between CCFRs are defined based on their support values and topological relationship as follows. Let the *successors* of a CCFR *r* be denoted as $succ(r)$ and be the set of CCFRs for which *r* is a prefix. Analogously, let the *predecessors* of a CCFR *r* be denoted as $pred(r)$ and be the set of CCFRs for which *r* is a suffix. Let the *free supply* of a CCFR *r* be $s(r) = sup(r) - \max_{r_i \in succ(r)} sup(r_i)$. Analogously, let the *free demand* of a CCFR *r* be $d(r) = sup(r) - \max_{r_i \in pred(r)} sup(r_i)$. Let $f(r)$ and $l(r)$ respectively denote first and last spatial element of a CCFR *r*. Let $conv(s)$ and $div(s)$ respectively denote the set of CCFRs that *converge to* and *diverge from* the spatial element *s*, i.e., $conv(s) = \{r_i : l(r_i) = s\}$ and $div(s) = \{r_i : f(r_i) = s\}$. A CCFR r_i *connects to* a CCFR r_j if and only if $l(r_i) = f(r_j)$.

Then, the initial probability of a CCFR r w.r.t. a set of trips TR is $\pi(r) = s(r)/|TR|$, and when r_i is connected to r_j, the transition probability from CCFR r_i to r_j is

$$\tau(i,j) = (1-\rho_{l(r_i)}) \sum_{r \in div(l(r_i))} \frac{d(r_j)}{d(r)}, \tag{1}$$

where $\rho_{l(r_i)}$ denotes the fraction of objects in the free supply of r_i that stop moving at r_i and is calculated as

$$\rho_{l(r_i)} = \max\Big(0, 1 - \big(\sum_{r \in div(l(r_i))} d(r) / \sum_{r \in conv(l(r_i))} s(r)\big)\Big). \tag{2}$$

When r_i is not connected to r_j then $\tau(i,j) = 0$. Note that the initial and transition probabilities effectively only distribute a free supply of a CCFR to connected CCFRs proportional to (1) and respecting (2) the free demand of those CCFRs.

3.3 Finding the k-MLMs

Given the domain projected CCFRs and the above movement model, a k-MLM problem with spatial and temporal predicates P_S and P_T is computed as follows. First, the *dow-hod* projected CCFR aggregates specified in P_T are retrieved and combined using the weighted CCFR mining method in (Bachmann et al., 2013). Second, a directed transition graph of CCFRs $G(E,V)$ is constructed as follows. Each CCFR r_i is added to G as a vertex v_i and each connected CCFR pair (r_i,r_j) is added to G as a directed edge e_{ij} with cost $-\log(\tau(i,j))$. In addition, an origin pseudo vertex v_o is created and a directed edge e_{oi} with cost $-\log(\pi(i))$ is created from v_o to a vertex v_i if the CCFR that corresponds to the vertex v_j intersects with the origin region $P_S.r_o$. Similarly, a destination pseudo vertex v_d is created and a directed edge e_{id} with cost 1 is created from a vertex v_i to v_d if the CCFR that corresponds to the vertex v_i intersects with the destination region $P_S.r_d$. Since in the so-constructed graph G the cost of a path from the origin pseudo vertex v_o to any other vertex that represents a CCFR is the negative log likelihood of the sequence of connected CCFRs $\langle r_1, r_2, \ldots, r_k \rangle$ that is represented by the path $\langle v_o, v_1, v_2, \ldots, v_k \rangle$, i.e., $cost(v_o, v_1, v_2, \ldots, v_k) = -\mathscr{L}(r_1, r_2, \ldots, r_k) = -\log(\pi(r_1)) + \sum_{i=1}^{k-1} -\log(\tau(i,i+1))$, the solution to the k-MLM problem is the set of k least cost paths from v_o to v_d in G. Since many sequences of CCFRs can generate a MLM, the likelihood of a MLM according to the movement model is the sum of the likelihoods of such sequences, which is calculated using a dynamic programming approach.

3.4 Finding Distinct *k*-MLMs

While the k-MLMs can theoretically be calculated with the method proposed by Yen (1971), the results may not be informative as several of the k-least cost paths in G may represent the same movement. Therefore, the following approach is proposed for the extraction of the *k-Most Likely Distinct Movements* (k-MLDMs). First, find the least cost path from v_o to v_d in G and return the corresponding sequence of CCFRs as the 1^{st}-MLDM. Second, find the set of CCFRs that are spatially contained by the 1^{st}-MLDM and increase the edge cost to vertices that represent these contained CCFRs by a user-defined *blocking factor* $f_b \geq 1$, resulting the modified graph $G^{1'}$. Subsequently, extract the 2^{nd}-MLDM as the least cost path in $G^{1'}$ and similarly iteratively extract the remaining $k-2$ MLDMs by 'blocking' the CCFRs that are contained by previously extracted MLDMs.

4 Demonstation

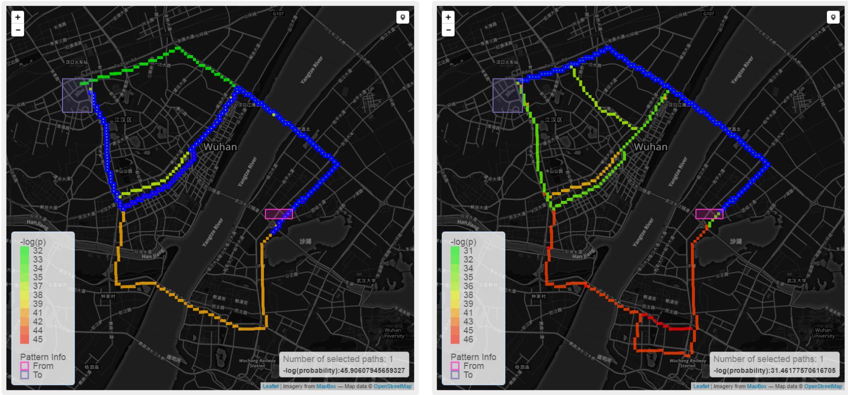
To demonstrate the proposed method a web application prototype is created and tested on a six day long (Mon, Tue, Thu, Fri, Sat, Sun) stream of raw GPS positions of around 11,000 taxis moving on the streets of Wuhan, China (Li et al., 2011). The data has been preprocessed into 2.26 million grid-based continuous trips as described in (Gidófalvi, 2015).

Incremental mining experiments reveal that the proposed method effectively compresses the infinite stream of trajectories to a finite storage space. In particular, at an absolute minimum support of 50 the processing of 2.26 million trips, partitioned into 145 *dow-hod* windows, results in a total of approx 4 million *dow-hod* projected CCFRs, which can be aggregated to 136 thousand global CCFRs, i.e, 5 times the number of CCFRs in an average *dow-hod* window. From these results one can estimate that the unbounded stream of movements of the taxis can be compressed into roughly 20 million CCFRs.

The two screenshots of the map interface in Figure 2 show the time-varying aspect of the k-MLDMs (morning vs afternoon) and the simple highlighting (blue) function that allows the interactive exploration of the k-MLDMs.

5 Conclusions and Future Work

The paper proposed a method that in an effective manner extracts complex, time-varying movement patterns (CCFRs) from a stream of moving object trajectories, regenerates likely movements based on these patterns, and facilitates the visual querying and explorations of these likely movements using a simple map interface.



(a) Morning (b) Afternoon

Fig. 2 Time-varying *k*-MLDMs.

Future work will consider 1) alternative models that further incorporate the topological relationship between CCFRs into transition probabilities, 2) empirical model validation, and 3) multi-modal movement model design.

References

Andrienko G, Andrienko N, Dykes J, Fabrikant S I and Wachowicz M, 2008, Geovisualization of Dynamics, Movement and Change: Key Issues and Developing Approaches in Visualization Research. *Information Visualization*, 7(3):173–180.

Andrienko G, Andrienko N, Demsar U, Dransch D, Dykes J, Fabrikant S I, Jern M, Kraak M-J, Schumann H and Tominski C, 2010, Space, Time and Visual Analytics. *IJGIS*, 24(10):1577–1600.

Bachmann A, Borgelt B and Gidófalvi G, 2013, Incremental Frequent Route Based Trajectory Prediction. *I*WCTS '13: Proceedings of the Sixth ACM SIGSPATIAL International Workshop on Computational Transportation Science, Orlando, FL, USA, 49–54.

Gidófalvi G, 2015, Scalable Selective Traffic Congestion Notification. *M*obiGIS '15: Proceedings of the 4th ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, Seattle, WA, USA, 10p.

Krogh B, Andersen O, Lewis-Kelham E, Pelekis N, Theodoridis Y and Torp K, 2013, Trajectory Based Traffic Analysis. *P*roceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Orlando, FL, USA, 536–539.

Krogh B, Pelekis N, Theodoridis Y and Torp K, 2014, Path-based Queries on Trajectory Data. *P*roceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Dallas, TX, USA, 341–350.

Li Q, Zhang T and Yu Y, 2011, Using Cloud Computing to Process Intensive Floating Car Data for Urban Traffic Surveillance. *IJGIS*, 25(8):1303–1322.

Yen J Y, 1971, Finding the K-Shortest Loopless Paths in a Network. *Management Science*, 17:712–716.