Proposal

November 5, 2017

Contents

1 Introduction

The goal of this research is to apply and develop a novel automatic traffic monitoring detection System based on available datasets in Hong Kong. More concretely, this research focuses on automatic incident detection, traffic state modeling and real-time routing based on probe car data, social media data and speed sensor data.

Accurate incident detection and route time estimation can be a valuable source of information. For routing algorithms, incident detection knowledge gives ground to a better estimate of travel times for different routes and improves the estimation of travel time.With the increasing use of cars, routing algorithms are a very important methodology to optimize traffic flow in larger cities.

2 Research Objectives

- Developing a system which could integrate social-media, GPS, speed panel and accident data to provide a high accuracy model of
  - traffic speed distributions in Hong Kong
  - traffic accidents locations together with a traffic anomaly indicator based on GPS, Speedpanel and social-media data
- Use the previously developed model for building an uncertainty-aware routing algorithm

3 Background of Research

3.1 Preprocessing and integration of the Traffic Data

Integration of traffic data with the aim of modeling traffic states falls into several steps. Independent of the number of data sources, preprocessing the traffic data (Transformation of raw data into a form which could be inserted into a chosen model) has very high importance. Failures in preprocessing the data can affect the end result a lot. Due to complexities of the transformations that are required to be performed on the data, it is very easy for mistakes to be made. Even if no logical mistakes are made, the geoprocessing algorithms and methodologies for GPS map matching and trajectory generation, if not chosen carefully, eliminate the possibility to build a high accuracy model. Thus state of the art trajectory based map matching algorithms will be implemented to match the low-frequency GPS data to roads [8] [1]. To complement the GPS data, high accuracy road sensor data is going to integrate to the input dataset. In addition to providing supplementary data, road sensors (RS), due to their higher accuracy can also contribute to assessing the likelihood of GPS observations. Because the RS don't have exactly same borders as the segments to which GPS data is matched to, specific methodology in regards to the RS spatial effect needs to be developed.

As the available road sensor data is based on averages, we are going to develop a methodology on how many GPS observations 1 speed panel observation equals to at a specific time and location Multiple additional considerations in terms of preprocessing, for example, whether filtering of the training dataset based on holidays, weekdays and such is necessary, will be carried out.

3.2 Model Building based on Traffic Data

After having developed preprocessing for the data, multiple state of the art traffic modeling techniques are going to be implemented and evaluated on 1 dataset. Traffic state modeling both for GPS and road sensor data has received broad coverage in recent literature and has focused mainly on 3 things: traffic flow estimation, traffic incident detection, and route time estimation. As methods that are proposed for GPS data have fewer assumptions, GPS data based models are going to be used in development. So far we have tested out a traffic incident detection method using Latent Dirichlet Allocation (LDA) on Hong Kong Data [6]. The incident detection results are satisfactory but the AUC values are 20 These models will be implemented to assess traffic speed distribution and the existence of accidents. The usefulness here lies in performing a systematical experimental study comparing the various state of the art methods both unsupervised and supervised with the same dataset and on a similar scale. Such a performance study has not been conducted before. The novelty comes from building up the model for 3 data sources: Social Media Data, Road Sensors data and GPS data. The experiments conducted so far using GPS and speed panel data have shown a variance of results. To solve this, each modeling result based on the quantity of the data in each location will be assessed using regular statistical reliability methods like p-value and trust intervals to give information how trustworthy each assessment is based on the underlying data. The initial experiments we have conducted so far appending speedpanel data have shown an improvement of 2 percent in regards to the AUC curves globally. In addition, the areas where speedpanel data is appended to GPS data perform better than areas without speedpanel data. However, due to the fact that the areas which have speedpanels are more presented in the data, there is lowering of results on roads that are away from the speedpanels. Thus, additional methodology will be developed to deal with this issue.

3.3 Development of Uncertainty Aware Routing Algorithms

Based on the results of the previous part we plan to build a probabilistic graph from the Hong Kong road system and investigate source to target queries(ST-queries) on top of it. Path queries on such graphs like finding the shortest travel time from the user location to the library are a very common execution sequence in nowadays map applications. Often these calculations are based on an average of collected observations of speeds and travel times in certain road links. The most well know algorithm for computing the path from start vertex to end vertex on a deterministic graph is the almost legendary A* algorithm [4] which calculates the route simply using the function $f(x) = h(x) + g(x)$ where $g(x)$ stands for the cost from start vertex to current vertex and h(x) stands for the heuristic value. These calculations, however, do not take into account the uncertainty of the given observations. To take account the probabilistic aspect we are now looking for the shortest-path distance that is the most likely to be observed when sampling a random graph from G. Taking account uncertainty in probabilistic graphs, however, is not an easy problem due to the intractable number of possible worlds an uncertain graph could generate. Calculation of ST-queries in uncertain graphs has been classified as P complicated problem [9]. One way to deal with this problem is using sampling [3] That means to sample a certain number of possible worlds from which it is possible to generate an approximation. Even with sampling, the space of execution can be extremely large so we plan to adapt solution proposed by Maniu et al. [7] to prune the size of possible worlds. The problem of creating uncertainty aware shortest path algorithms in traffic has been focused before in Hua et al. [5] which proposed multiple graph-based algorithms that estimated travel times in graphs within a certain trust interval and based its calculations on edges which consisted probability distributions over travel times. This method also takes into account the dependence of edges in the graph.This research will extend this method by previously pruning out possible worlds. In addition, as the input generation for the probabilistic graph is also under our control, we can devise overall the best features and distributions for this step.

3.4 Prototype for the Intelligent Traffic Monitoring System

After proposing and theoretically developing the model we will build a prototype system which can efficiently manage ST-queries calculation and traffic state classification. This prototype will be built using a distributed environment taking advantage as much as the possible state of the art parallel computing software like Spark, Hadoop, and Genomes. The prototype is going to be used to perform tests using the available road sensor and speed panel datasets. In addition, to visualize the results in real time, we are going to use frameworks for OSM data [2].

4 Research Plan and Methodology

The research performed is going to be developmental and experimental. Existing implementations of different traffic state models are going to be tested out and extended for additional data. A new routing algorithm is going to be devised on based on the generated probabilistic graph on top of Hong Kong road system.

5 Expected Outcomes and Significance

Currently, It can be said, that methods divide into groups based on multiple attributes. There exists a split between works which use GPS Probe Car Data(PCB) for building their models and works which base their analysis on Road Sensors (RS). By Road Sensors it is meant sensors installed aside a road, which collects data mainly about the speed and volume of the traffic. No work has, however, been focused on integrating these data sources together. This is without a reason because traffic monitoring and route planning are most used and most important in large cities where both data sources GPS and RS data is often available. GPS data here can complement the RS data by having a very broad reach whereas RS can prove helpful in assessing the GPS data and providing a larger input for the general training of the model which is important to avoid overfitting. In addition, a third type of data source by the face of Social Media is going to be incorporated into the model. In addition, a comparative incident detection performance study has not been performed on 1 unified dataset with 1 specific methodology. Such research can be a valuable addition to the existing works.

The result of assessing traffic is still somewhat inaccurate due to the uncertainty of the data. This fact of uncertainty has not been widely researched in generating ST-queries for urban traffic data. Based on the author's knowledge only 1 paper has been written by Hua et al. [5]. We will devise a new routing algorithm which also takes into account the pruning of possible worlds.

6 Study schedule

First year We plan to perform an experimental study comparing multiple different states of the art traffic modeling and incident detection algorithms on Hong Kong data.Through this, we can achieve a clear perception of existing incident detection methodologies used and find out each different methods weaknesses and strengths on 1 dataset. This comparative study also allows me to get a good sense of features used in incident detection. The incident detection system I have developed during my master's thesis for preprocessing traffic data allows me to implement new models quickly. Incorporating multiple datasets is going to provide 1-2 publications in top conferences and journals. The performance study will provide 1 additional publication.

Second and Third year we plan to choose one of the developed models for ground truth data and fully develop a uncertainty aware route planning algorithm on top of Hong Kong roads. This development is going to provide additional 2 publications in top journals and conferences.

Fourth year we are going to fully develop a traffic monitoring system based on the developed methods. As traffic data and analysis is a very relevant topic which gets a lot of interest from the general public, we are going to take advantage of the fact that geoprocessing projects provide a lot of opportunities for visualization. Through visualization, it's likely easier to get the attention of the media and gain publicity.

References

[1] B. Y. Chen, H. Yuan, Q. Li, W. H. Lam, S.-L. Shaw, and K. Yan. Map-matching algorithm for large-scale low-frequency floating car data. *International Journal of Geographical Information Science*, 28(1):22–38, 2014.

[2] O. S. M. Corporation. Osm frameworks, 2017.

[3] G. S. Fishman. A monte carlo sampling plan for estimating network reliability. *Operations Research*, 34(4):581–594, 1986.

[4] P. E. Hart, N. J. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

[5] M. Hua and J. Pei. Probabilistic path queries in road networks: traffic uncertainty aware path selection. In *Proceedings of the 13th International Conference on Extending Database Technology*, pages 347–358. ACM, 2010.

[6] A. Kinoshita, A. Takasu, and J. Adachi. Real-time traffic incident detection using a probabilistic topic model. *Information Systems*, 54:169–188, 2015.

[7] S. Maniu, R. Cheng, and P. Senellart. An indexing framework for queries on probabilistic graphs. *ACM Transactions on Database Systems (TODS)*, 42(2):13, 2017.

[8] M. Quddus and S. Washington. Shortest path and vehicle trajectory aided map-matching for low frequency gps data. *Transportation Research Part C: Emerging Technologies*, 55:328–339, 2015.

[9] L. G. Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3):410–421, 1979.