Bus Routes Design and Optimization via Taxi Data Analytics

Seong Ping Chuah, Huayu Wu, Yu Lu, Liang Yu
Institute for Infocomm Research
1 Fusionopolis Way, Singapore 138632
{chuahsp,huwu,luyu,yul}@i2r.a-star.edu.sg

Stephane Bressan
School of Computing,
National University of Singapore
13 Computing Drive, Singapore 117417
steph@nus.edu.sg

ABSTRACT

Public bus services are often planned in the context of urban planning. For a city with efficient and extensive network of public transportation system like Singapore, enhancing the existing coverage of bus service to meet the dynamic mobility needs of the population requires data mining approach. Specifically, frequent taxi rides between two locations at a period of time may suggest possible poor coverage of public transport service, if not lacking of the public transport service. In this paper, we describe a proof of concept effort to discover this weakness and its improvement in public transportation system via mining of taxi ride dataset. We cluster taxi rides dataset to determine some popular taxi rides in Singapore. From the clustered taxi rides, we filter and select only the clusters whose commuting via existing public transport are tortuous if not unreachable door-to-door. Based on the discovered travel pattern, we propose new bus routes that serve the passengers of these clusters. We formulate the bus planning problem as an optimization of directed cycle graph, and present it's preliminary solution and results. We showcase our idea in the case of Singapore.

Keywords

Taxi Rides, Clustering, Bus Route Design & Optimization

1. INTRODUCTION

Singapore is well known for its high quality and efficient public transport service. Millions of people rely on its high quality, efficient, and extensive network of public bus and mass rapid transit (MRT) to move around everyday. Despite its efficiency, there are 28,286 strong of taxi fleet as of March 2016 [1], and 519,645 private cars on road in 2015 [2]. The figure is compared to just 17,740 buses (all types) in 2015 [2].

[1] www.lta.gov.sg/content/dam/ltaweb/corp/PublicationsResearch/files/FactsandFigures/taxi_info_2016.pdf

[2] www.lta.gov.sg/content/dam/ltaweb/corp/PublicationsResearch/files/FactsandFigures/MVP01-1_MVP_by_type.pdf

CIKM '16, October 24–28, 2016, Indianapolis, IN, USA.

DOI: http://dx.doi.org/10.1145/2983323.2983378

Public bus service can be extended to offer more efficient solutions to the mobility needs of the population. Traditionally, public bus services are often planned in the context of urban planning [2] and operation research [3]. For a city with efficient and extensive network of public transportation system, enhancing the existing coverage of bus service requires data mining approach. Specifically, frequent taxi rides between two locations at a particular period of time may suggest possible poor/lacking of bus route services in the existing transportation network at that particular period of time. An extreme situation is usually found after midnight, where MRT lines and bus services cease to operate while taxis becomes the only public transport. Currently, Singapore has 14 routes of night bus operating from 12am to 2am of Friday, Saturday, and eve of public holiday. However, these bus routes are designated to transport party-goers from downtown city center returning to various suburban residential areas before 2am rather than the transportation needs of the nightlife-goers throughout midnight.

On the other hand, improving the efficiency of the existing public transportation system of Singapore is non-trivial, given its current level of efficiency. We seek to improve the efficiency of public transportation system by determining new bus route solutions for a specific period of time. Travel patterns of these passengers are not available in the trip records of the public transportation system. A natural way is to refer to the statistics of taxi rides. Mining the taxi rides data provides insights to the travel patterns of not only these taxi takers, but also potentially the private car owners whose lifestyle and consumer behaviours are likely similar.

In this paper, we showcase a proof of concept effort in improving the public transportation network via mining of taxi ride dataset. We mine the most popular origin-destiny of taxi ride within a specific period of time from a dataset of taxi rides. Based on the clusters of trajectory, we determine the popular origins and destinies where people would take taxi. We filter out the clusters whose journeys can be commuted easily (within a single trip ride) by the existing public transportation system. We study the remaining clusters whose commuting via public transports are tortuous if not unreachable door-to-door. Based on these taxi ride clusters and travel patterns, we propose new bus routes for the specific time windows that serve the passengers of these clusters by minimizing the distance the buses travel without passengers, while ensuring quality of service such as travelling time and bus load.

The paper is organized as follows. We discuss the mining of taxi rides data in Section 2. In Section 3, we discuss the

planning problem of new bus routes, and formulate its optimization problem and solution. We showcase a case study in Section 4. The paper is concluded in Section 5

2. CLUSTERING OF TAXI RIDES

Taxi service provides the most flexible transportation service in a city. In a modern city with extensive coverage of public buses and MRT, frequent taxi rides from one location to another indicate high demand of travelling on the route that is not covered by the public transportations. In this section, we discuss the mining of these frequent taxi rides to discover potential new bus routes.

2.1 DBSCAN Clustering

We aim to discover these travel patterns of frequent taxi ride from the taxi ride data, which contains GPS data and time stamp of the origin and destiny of each taxi ride. We perform clustering on the data of taxi rides within a period of time. Taxi pick-up and drop-off locations are usually along roadsides, their geographical distributions are non-globular. Thus, we deploy density-based method, DBSCAN for clustering of taxi rides. DBSCAN [4] has the advantages of robustness to noise and capability to handle clusters in arbitrary shape of road. As each taxi ride consists of a pair of pick-up and a drop-off point, each data sample input to DBSCAN is a four-dimensional coordinates (coordinate-X and coordinate-Y for both points).

2.2 Clusters Filtering

DBSCAN clustering returns clusters of frequent taxi rides. However, despite their high number of ridership, not every cluster suggests potential new routes of bus service. Some clusters should be filtered out based on several factors as in the following discussions. After filtering out these unsuitable clusters, clusters with the higher number of point (taxi ride) are selected for new bus route planning and optimization.

2.2.1 Availability of Public Transport

From the clustering, we may get many clusters of taxi rides which is covered by the existing public buses or MRT service. These clusters represent the group of passenger who is willing to pay a bit more for easier, faster, and more convenient way to reach home. These clusters are often found in suburban residential areas of Singapore. A suburban residential area usually has an MRT station, a bus interchange and shopping malls co-located as the hub to the residential area. People take a short range taxi ride from the hub to their apartment, despite frequent community buses plying the neighbourhood of the residential area from the hub. For these clusters, there is no new bus route to be proposed.

2.2.2 Airport Traffic

Many people travel by air for business and vacation purposes in Singapore. From the clustering result, we may get many clusters which have the airport either its origin or destiny. While there are buses and MRT service connecting the airport, taxi is still a popular choice for passenger with luggages. Departure passengers take taxi to reduce the risk of running late, while the arrival passengers take taxi to reach home/hotel faster after tiring flight. Thus, these clusters do not suggest any lack of coverage of public bus plying to and from the airport.

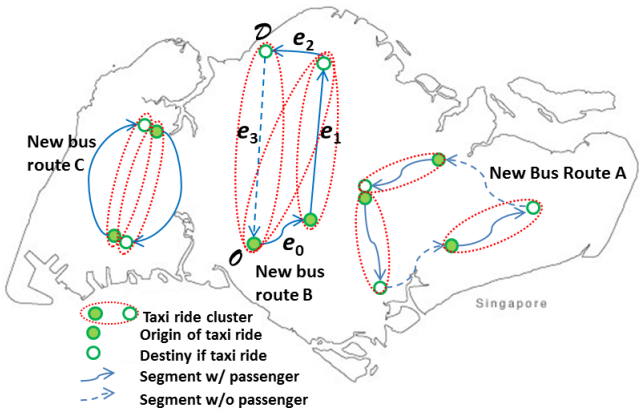

Figure 1: Linking the origins and destinies of all clusters with directed cycle graphs.

3. NEW BUS ROUTES OPTIMIZATION

Based on the clusters of taxi rides obtained from the clustering, we plan new bus routes for the travelling needs of these clusters. Given sets of taxi rides with an origin and a destiny for each cluster, we optimize the bus routes to transport the passengers of these clusters in the most efficient way. In this section, we present the routing problem of the new buses, and some preliminary solution to the problem.

3.1 The Bus Routing Problem

The bus routing problem for several clusters is illustrated in Fig.1. The bus routing problem bears some resemblance to the travelling salesman problem (TSP)[1] where the salesman needs to visit all cities in a single trip. But unlike TSP where the salesman is free to visit the cities in any order, the new bus should visit the origin before the destiny of each cluster. That limits the search space of the next location the bus should visit in the combinatorial optimization. Besides, the bus routing problem is not limited to just one bus route to serve all clusters. To sustain the dispatching of buses over a long period of time, a new bus route that passes through the cluster(s) should form a directed cycle graph.

Fig.1 shows three types of new bus routes identified based on the taxi ride clusters. In practice, a new bus route can be a combination of these three types. An origin (green filled circle) and a destiny (green unfilled circle) form a taxi ride cluster (circled by red dotted line). Solid blue lines are the segments of bus route that carries passenger from their origins to destinies, while dash blue lines are the segments of bus route whereby the bus travel to the next taxi ride origin to pick up the passenger.

It can be seen that all new bus routes in Fig.1 form directed cycles. To reduce operating cost, it is desired that the blue dash line be minimized, since empty bus receives no bus fare. Thus, New Bus Route A is the least preferable case where significant fraction of the bus route are travelled without passenger load. Conversely, New Bus Route C is the most preferable case where shuttle buses plying both locations serve two clusters with opposite travelling directions. In New Bus Route B, two clusters (generalized to N clusters) are merged into a route travelling to a same direction before returning to the origin with empty load. Merging more clusters into New Bus Route B results in less operat-

ing cost, but degrade quality of service for some passengers. It discourages them from switching to public bus from the existing taxi rides. Under New Bus Route B, the path from origin $\mathcal{O}$ to destiny $\mathcal{D}$ is actually $\{e_0 \rightarrow e_1 \rightarrow e_2\}$ rather than the direct opposite direction of e_3. Thus, the travelling time for this cluster ($\mathcal{O}$-$\mathcal{D}$ pair) under New Bus Route B is longer than that of separate bus route for both clusters. Besides, buses travelling on e_1 may be overcrowded due to the merging of two clusters in a service route.

3.2 Optimization Formulation

We formulate the bus routing problem as an optimization problem of directed cycle graphs. Let $\mathcal{C}_n$ be the taxi ride cluster in the set $n \in \mathcal{N}$ to be served by new bus routes. Each cluster consists of a pair of origin ($\mathcal{C}_n.\mathcal{O} \in \mathcal{V}$) and destiny ($\mathcal{C}_n.\mathcal{D} \in \mathcal{V}$) serving as vertices in set $\mathcal{V}$. These clusters are served by set of proposed bus routes $\mathcal{R}_m, m \in \mathcal{M}$ which is directed cycle graph $\mathcal{R}_m = \{e_m(\mathcal{C}_n.\mathcal{O}, *), \cdots, e_m(*, \mathcal{C}_n.\mathcal{D}), \cdots, \tilde{e}_m(*, \mathcal{C}_n.\mathcal{O})\}$ comprising a series of directed edges e_m with passenger load, and directed edges with empty load $\tilde{e}_m$. $e(a, b)$ is an edge that travels from vertex a to vertex b. $\{e_m(\mathcal{C}_n.\mathcal{O}, *), \cdots, e_m(*, \mathcal{C}_n.\mathcal{D})\}$ represents that bus route m picks up passengers of $\mathcal{C}_n$ at $\mathcal{C}_n.\mathcal{O}$, and drops them at $\mathcal{C}_n.\mathcal{D}$ after passing through some other vertices/edges. Similar notation for $\tilde{e}_m$.

We optimize the bus routes such that the travel distances of empty buses $\tilde{e}_m(*, *).dist$ are minimized while all clusters are served with some quality of service. The optimization problem can be formulated as

$$\min_{\mathcal{R}_m} \quad \sum_{m \in \mathcal{M}} \sum_{dist} \tilde{e}_m(*, *).dist \tag{1}$$

$$\text{s.t.} \quad \mathcal{R}_m.arg(\mathcal{C}_n.\mathcal{O}) < \mathcal{R}_m.arg(\mathcal{C}_n.\mathcal{D}) \ \ \forall\, m, n \tag{2}$$

$$\mathcal{R}_m = \{e_m(\mathcal{C}_n.\mathcal{O}, *), \cdots, \tilde{e}_m(*, \mathcal{C}_n.\mathcal{O})\} \ \ \forall\, m \tag{3}$$

$$\mathcal{R}_m.time(\mathcal{C}_n) \leq \alpha \cdot \mathcal{C}_n.time \ \ \forall\, m, n \tag{4}$$

$$e_m(*, *).load \leq \text{max\_load} \ \ \forall\, m \tag{5}$$

where (2) ensures that the origins of all clusters are visited before its destinies in all bus routes; (3) ensures that all bus routes are directed cycle graphs; (4) and (5) ensure the quality of bus service. (4) requires that the travelling time for all clusters in all bus routes should be less than a factor (α) of the mean travelling times ($\mathcal{C}_n.time$) of taxi rides of their corresponding cluster. $\alpha > 1$ is acceptable since taxi rides are usually faster. (5) ensures no overcrowding at all edges in all bus routes.

3.3 Solution Method

The formulated problem (1)-(5) is a combinatorial optimization which is NP-hard in general. To determine new bus routes, we develop a heuristic method to obtain a preliminary solution. Specifically, we aim to merge as many clusters as possible subject to the constraints of (4)-(5). We exploit constraint (2) to reduce the search space of the next vertex to visit, where candidates of the next vertex comprise

1. the departing vertex of the route, ie the origin of the first cluster it visited
2. destinies of all clusters visited by the route before
3. the origins of clusters not served by any bus route

Point 1 ensures that the route ends at the departing vertex, thus forming a cycle graph. We deploy a greedy approach to select the shortest distance among the candidate vertices which satisfy the load and time constraints in (4)-(5). Both constraints impose a limit to the merging of clusters to avoid overcrowding of buses and long travelling time. When no origin of other clusters can be admitted into the candidate vertices (due to constraints (4)-(5)), and destinies of all clusters has been visited, the route end in the departing vertex of the route. The bus routing algorithm is run to determine another new bus route until all taxi ride clusters are served by bus routes. Algorithm 1 summarizes all.

Algorithm 1 Algorithm for discovering new bus routes

```
Initialize m = 0
while |N| > 0 do
  Initialize a new route R_m
  x_cur = C_n.O // Initialize an origin as the current vertex
  Initialize V = C_n.D ⋃ {C_n.O, ∀ n ∈ N} // All origins in N + destiny of the current vertex
  while |V| > 0 do
    x* = arg min_{x∈V} e_m(x_cur, x).dist s.t. (4)-(5)
    x_cur = x*
    if (x* == O) // if the next vertex is an origin then
      V = V ⋃ arg(x*).D // include its destiny in the set of candidate vertices
    else
      N = N \ arg(x*) // the cluster is served, so removed the corresponding cluster from N
    end if
  end while
  m = m + 1
end while
```

4. CASE STUDY

We showcase a case study to identify new bus routes using historical dataset of taxi rides in Singapore. The entire taxi rides dataset comprises roughly 7.1 millions taxi rides recorded by 15028 taxis in Singapore. The dataset contains the GPS records and the time stamps of pick-up and drop-off locations of each taxi ride. We first clean the dataset by removing negligible number (around 20 out of roughly 7.1 millions records) of erroneous records of GPS data.

In metropolitan Singapore, travel pattern of its population depends largely on the work week. In particular, Friday evening is among the busiest periods on Singapore's roads. We mine the taxi rides dataset and discover that Friday evening, from 5pm till 12am midnight to be the period of time where clusters of taxi rides link business areas to shopping/entertainments/restaurants/chill out areas not directly accessible via public transportation network. DBSCAN clustering of the taxi rides and filtering of the clusters are as described in Section 2.

We obtain $|\mathcal{N}| = 8$ suitable clusters of most popular taxi rides. Fig.2 and Fig.3 show the locations of these clusters. Table 1 shows the origins and destinies, and the traffic volume of each cluster. Using Algorithm 1 where $\alpha = 2$ and max_load = 70, we obtain $|\mathcal{M}| = 3$ new bus routes where each route serves the following clusters: $\mathcal{R}_1 = \{\mathcal{C}_4, \mathcal{C}_6, \mathcal{C}_7, \mathcal{C}_8\}$, $\mathcal{R}_2 = \{\mathcal{C}_1, \mathcal{C}_3, \mathcal{C}_5\}$, $\mathcal{R}_3 = \{\mathcal{C}_2\}$. Paths of the bus routes are given in Table 2 where

$$\tilde{d}_m \triangleq \sum_{dist} \tilde{e}_m(*, *).dist \ , \ \ d_m \triangleq \tilde{d}_m + \sum_{dist} e_m(*, *).dist \tag{6}$$

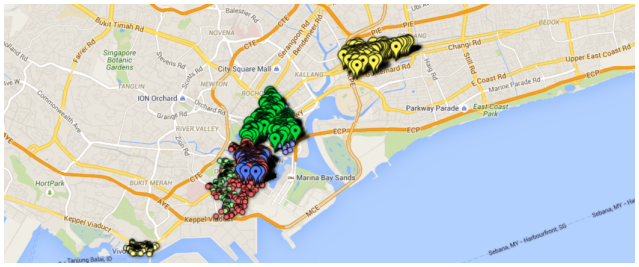

Figure 2: Clusters of taxi ride origins

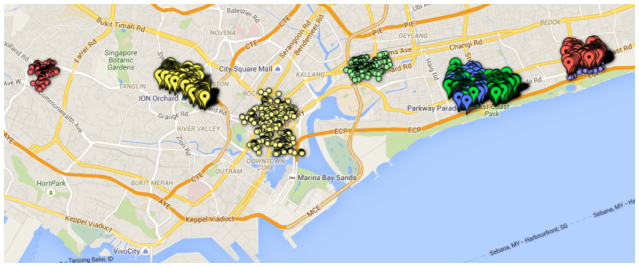

Figure 3: Clusters of taxi ride destinies

All these routes takes long time to commute and transit via existing public transport, are not reachable via MRT or any single bus ride. From the table, $\mathcal{R}_2$ is the most efficient route as it has the shortest distance (and fraction) of path without passenger load, while $\mathcal{R}_3$ has the highest fraction of path without passenger load as it only serves $\mathcal{C}_2$. $\mathcal{R}_1$ has the longest distance of path without passenger load. Table 3 compares the mean travelling time of taking taxi ($\mathcal{C}_n.time$) with expected travelling time taking the new bus routes ($\mathcal{R}_m.time(\mathcal{C}_n)$). We deploy the API of Singapore Journey Planner [5] developed at Institute for Infocomm Research to obtain the route and expected travelling time. The expected travelling times for all clusters under the proposed new bus routes are within twice of the mean travelling time of taxi rides. Note that taking bus is usually much cheaper than taxi in Singapore. Fare analyses of both transportation modes are left out due to page limit.

5. CONCLUSIONS

We present in this paper a proof of concept in planning new routes of public bus service via data mining approach. Through the mining of taxi dataset, we identify some leaks of coverage of public transportation network at a specific time windows within a week. Based on the travel patterns of taxi rides clusters, we show that the efficiency and quality of service of the public transport can be improved by proposing new feasible bus routes for a specific time windows. We illustrate our proof of concept in a case study of Singapore. For future works, passengers' travel patterns in public transports can be mined to jointly optimize the routings of existing and newly proposed bus routes. Besides GPS data, fare data of passengers of both transport modes can be factored in to improve the competitiveness of public transports. Finally, the framework can be extended to cater for dynamic demand of mobility needs in a work week.

Table 1: Clusters of taxi rides

Cluster	Origin	Destiny	Vol. (taxi/hour)
$\mathcal{C}_1$	Harbour Front	Civic District	13.5
$\mathcal{C}_2$	CBD	Holland Village	9.7
$\mathcal{C}_3$	Geylang	Orchard Rd.	7.6
$\mathcal{C}_4$	Civic District	Marine Parade	7.6
$\mathcal{C}_5$	Chinatown	Geyland	6.7
$\mathcal{C}_6$	Civic District	East Coast	5.7
$\mathcal{C}_7$	CBD	East Coast	4.5
$\mathcal{C}_8$	CBD	Marine Parade	4.0

Table 2: Proposed new bus routes

Route	Path	$\tilde{d}_m$ / d_m (km)
$\mathcal{R}_1$	CBD → Civic District → Marine Parade → East Coast → CBD	13.9 / 31.8
$\mathcal{R}_2$	Harbour Front → Chinatown → Civic District → Geylang → Orchard Rd. → Harbour Front	7.5 / 27.0
$\mathcal{R}_3$	CBD → Holland Village → CBD	8.7 / 19.3

Table 3: Comparison of travelling times

Cluster	$\mathcal{C}_n.time$ (min)	$\mathcal{R}_m.time(\mathcal{C}_n)$ (min)
$\mathcal{C}_1$	17.8	22
$\mathcal{C}_2$	18.7	27
$\mathcal{C}_3$	16.1	25
$\mathcal{C}_4$	19.1	33
$\mathcal{C}_5$	16.6	24
$\mathcal{C}_6$	18.6	24
$\mathcal{C}_7$	18.9	37
$\mathcal{C}_8$	19.4	28

6. ACKNOWLEDGMENTS

This research is partially supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. NUS Ref: R-702-005-101-281.

7. REFERENCES

[1] D. Applegate, R. Bixby, V. Chvátal, and W. Cook. *The Traveling Salesman Problem: A Computational Study.* Princeton Series in Applied Mathematics. Princeton University Press, 2011.

[2] C. Daraio, M. Diana, F. D. Costa, C. Leporelli, G. Matteucci, and A. Nastasi. Efficiency and effectiveness in the urban public transport sector: A critical review with directions for future research. *Eur. J. of Operational Research*, 248(1):1 – 20, 2016.

[3] V. Schmid. Hybrid large neighborhood search for the bus rapid transit route design problem. *Eur. J. of Operational Research*, 238(2):427 – 437, 2014.

[4] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining, (First Edition).* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[5] L. Yu, D. Shao, and H. Wu. Next generation of journey planner in a smart city. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 422–429, Nov 2015.