ABSTRACT

Customized bus (CB) services that operate between residential areas and work places during rush hours have surfaced in many cities to deal with traffic congestion problem. One major challenge of CB services is to design services based on user demands. To solve this problem, we extract commute patterns from large-scale taxi GPS data for CB service area design. We present a novel clustering algorithm that embeds origin and destination data to a low dimensional feature space, such that the resulting patterns satisfy both the needs of commuters and operators. The algorithm has been tested on both synthesized and real-world data.

1 INTRODUCTION

Traffic congestion is a severe problem worldwide, with peak-hour congestion being the most frustrated issue for daily commuters. In 2014, approximately 55% delay is caused by peak-hour congestion which costs U.S. commuters extra 42-hour of travel time and 19-gallon of fuel waste averagely[10].

On-demand bus service companies like BRIDJ and GoGoBus allow users to submit their intended trips and optimize pick-ups, drop-offs and routes. Such services provide better riding experience than traditional public transport. Compared with car-sharing services, CB that carries more people than shared cars would help reduce peak-hour congestion even further[5]. Therefore, CB services can be beneficial to both commuters and the city traffic authorities.

However, current design process for customized bus service areas still have some shortcomings. First, candidate routes can only be submitted manually by users; second, changing routes frequently is inconvenient for both operators and riders. Therefore, a data-driven approach that extracts stable, salient commute patterns is needed.

Due to the ubiquity of GPS devices, large-scale taxi GPS datasets are frequently used in mobility pattern mining, which can reliably reflect the global repetitive commute demands[1]. In previous work of designing origination (O) and destination (D) service areas, [2] only considered hotness of pick-ups and drop-offs separately: [8] put OD relationship into consideration but broke it into segment-and-pair two-phase process, during which information can be lost. [7] used T-DBSCAN clustering method, which considered OD relationships in one phase but has many design parameters to choose.

In this paper, we propose a new method of customized bus service area design. The main contributions are:

1. We propose a new clustering method to embed high dimensional data into low dimensional feature for clustering, which is a generalized method for high dimensional data clustering.
2. We design a customized bus service area selection method that optimize both user convenience and resource cost.

2 ALGORITHM

Our method is designed to satisfy the needs of both users and service operators. Users want to minimize their walking distances to bus stops; operators, on the other hand, want to maximize number of people on the same bus so that people originate from the same area should end around the same place.

Formally, we define the service area design problem as follows: Given the origins X and destinations Y of N taxi trips during a given time-of-day, we want to find feature functions \( f : X \rightarrow \mathbb{R} \) and \( g : Y \rightarrow \mathbb{R} \), such that

1. The intra-cluster weighted distance \( X \) and \( Y \) are minimized

2. The correlation between \( f(X) \) and \( g(Y) \) is maximized

In objective (1), 2-D data \( X \) and \( Y \) are embedded into 1-D feature space that optimize both user convenience and resource cost.

Objective (2) uses HGIR maximal correlation[9], the only dependence measure that satisfies Rényi’s seven postulates, to quantify the relationship between OD pairs.

This formulation (1), (2) have several benefits. First, correlation term provides a tradeoff between walking distance and OD correlation. Second, \( f(\cdot) \) and \( g(\cdot) \) are real-valued and solving continuous optimization problem is easier than integer ones. Third, \( f(\cdot) \) and \( g(\cdot) \) can be treated as 1-D embedded feature for further clustering without the need to specify crucial parameters explicitly.

Note that this formulation has a trivial solution, where clusters can be formed by individual points. We solve it by adding a data weighted regularization term [6] to reduce number of clusters.

Combining (1), (2) and (3) with parameters \( \gamma \) for correlation trade-off and \( \lambda \) for regularization trade-off. The final optimization problem is shown below. Although it is a non-convex optimization problem, it can be efficiently solved using projected gradient descent to converge.

\[
\begin{align*}
\min_{f,g} & \frac{1}{2} \sum_{i,j} e^{-||X_i - X_j||} ||f_i - f_j||^2 + \frac{1}{2} \sum_{i,j} e^{-||Y_i - Y_j||} ||g_i - g_j||^2 \\
\text{s.t.} & \quad \mathbb{E}[f(X)g(Y)] = \gamma, \mathbb{E}[f^2(X)] = \mathbb{E}[g^2(Y)] = 1
\end{align*}
\]
Table 1: Clustering Similarity Index of Our Method Compared with 4d-kmeans and DBSCAN

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4d-kmeans</td>
<td>0.58 ± 0.12</td>
<td>0.73 ± 0.10</td>
<td>0.59 ± 0.09</td>
</tr>
<tr>
<td>DBSCAN[13]</td>
<td>0.76 ± 0.09</td>
<td>0.84 ± 0.09</td>
<td>0.70 ± 0.07</td>
</tr>
</tbody>
</table>

Figure 1: Synthesized X-shape Data Result. (a) Histogram of 1-D embedded features f(origination) and g(destination) (b) Four identified X-shape routes

Figure 2: Effect of $\lambda$ on Number of O/D Clusters on Synthesized Data

3 RESULTS

Synthesized Data Result. Synthesized data $\{X_i, Y_j\}_{i=1}^{200}$ contains 200 trips sampled from 4 unique routes from two origins (o1, o2) to two destinations (d1, d2) with added gaussian noise. Figure 1.a shows 4 peaks of embedded 1-D feature $f(X)$ and $g(Y)$ with $\gamma = 1$ and $\lambda = 130$. Further clustering these features, we obtain result in figure 1.b, which perfectly recovers all 4 ground truth routes. We also validate the regularization effect of $\lambda$ in Figure 2: number of clusters decreases as $\lambda$ increases. $\lambda \in [60, 130]$ gives ground truth result, which is the longest interval of $\lambda$.

Real Data Result. NYC Taxi data[11] trips with distance $d \in [7km, 10km]$ during 17:00-19:00, May 11th, May 22nd, 2015 are considered. Setting parameters $\gamma = 1$ and $\lambda = 30$, our algorithm found 12 origin and 15 destination service areas with 18 routes. In Figure 3, six major commute patterns are shown, which are from Brooklyn downtown to Brooklyn residential areas (Route3) or to Manhattan midtown (Route4), from upper Manhattan to lower Manhattan (Route2) or to Bronx (Route5), from Harlem District (Route6) and Long Island City (Route1) to LaGuardia Airport. Our result is consistent with the typical commute patterns of New York City residents.

Comparisons. We compare our method with classic clustering methods, 4D-kmeans and DBSCAN[13]. Similarities between our method and baseline methods are computed using Adjusted Rand Index[4], Hubert’s Index[3] and Adjusted Mutual Information[12] under different parameter settings($k \in [2, 10]$ for kmeans, $\epsilon \in [0.1, 0.2]$ and $n_{min} \in [6, 16]$ for DBSCAN) shown as ‘mean±std’ in Table 1. DBSCAN selects salient areas with large densities but lack good coverage while kmeans clusters all the points but have lower density. Our algorithm have similarities to both methods (all the indexes are close to 1), which combines coverage and density.

Figure 3: NYC Data Result - 6 Major Routes (17:00-19:00, May 11th to May 22nd, 2015)

4 CONCLUSIONS

This paper proposes a new method for CB service area design, which considers O/D connections and embeds original data into 1-D feature space so that intrinsic nature of data can be obtained for further clustering. It also introduces the concept of maximal correlation to quantize the system performance, enabling the operator to trade-off between user satisfaction and operation cost. In the future, we plan to design detailed bus routes within service areas.

ACKNOWLEDGMENTS

The author would like to thank Dr. Yang Li for discussions and advisors Prof. Zhang and Prof. Shao-lun Huang for guidance.

REFERENCES