Identifying Tourists from Public Transport Commuters

Mingqiang Xue #, Huayu Wu #, Wei Chen #, Wee Siong Ng #, Gin Howe Goh *
#Institute for Infocomm Research, A*STAR, Singapore
{xuem, huwu, chenwei, wsng}@i2r.a-star.edu.sg
*Innovation & InfoComm Group, Land Transport Authority of Singapore
gin_howe_goh@lta.gov.sg

ABSTRACT

Tourism industry has become a key economic driver for Singapore. Understanding the behaviors of tourists is very important for the government and private sectors, e.g., restaurants, hotels and advertising companies, to improve their existing services or create new business opportunities. In this joint work with Singapore's Land Transport Authority (LTA), we innovatively apply machine learning techniques to identity the tourists among public commuters using the public transportation data provided by LTA. On successful identification, the travelling patterns of tourists are then revealed and thus allow further analyses to be carried out such as on their favorite destinations, region of stay, etc. Technically, we model the tourists identification as a classification problem, and design an iterative learning algorithm to perform inference with limited prior knowledge and labeled data. We show the superiority of our algorithm with performance evaluation and comparison with other state-of-the-art learning algorithms. Further, we build an interactive web-based system for answering queries regarding the moving patterns of the tourists, which can be used by stakeholders to gain insight into tourists' travelling behaviors in Singapore.

Categories and Subject Descriptors

I.6.4 [**Computing Methodologies**]: Simulation and Modeling; K.4.2 [**Computers and Society**]: Social Issues; H.4.0 [**Information Systems Applications**]: General

Keywords

EZ-link; tourists; data analytics; public transport

1. INTRODUCTION

1.1 Context

Using data analytics to improve the quality of various services and cares in society has attracted increasing attention from governments and private sectors in many countries. The Singapore government started promoting data analytics in different domains a few years back. Plenty of data from government agencies have been published via data.gov.sg to encourage innovative applications and R&D attempts from research institutes and private companies. On the other hand, for those data that cannot be publicly shared due to various reasons, e.g., containing sensitive information, relevant agencies often choose to work with the government-based research institutes to discover useful knowledge from the data.

Since 2013, the Land Transport Authority (LTA) of Singapore initiated a couple of projects with the Institute for Infocomm Research (I^2R) under Singapore's Agency for Science, Technology and Research, aiming to analyze public (e.g., MRT[1] and bus) and private (e.g., taxi and private car) transport data to come out with insights on areas to improve their existing services, and also to leverage on transport data analysis to get insights on social activities that would be helpful to other sectors for the benefit of Singapore's economy and development.

1.2 Motivation and Challenges

As a famous tourist city in Southeast Asia, Singapore is attracting more than 10 million foreign visitors yearly, with 23 billion Singapore dollar tourism receipts in 2012. The government is continuously making efforts to promote the tourism industry and build it into a key economic driver for Singapore.

The living places, the visited places and the travel patterns of tourists are important information for relevant public and private sectors to design and improve their services in the tourism industry. For example, souvenir store can be opened at places where tourists mostly visit, recommendations and special packages can be made if two or more places are often commonly visited, additional bus services can be introduced for popular routes seasonally, advertisement can be targeted at MRT and bus stations where tourists make transit, and etc. Such tourist information is traditionally collected by surveys and the result is reported annually by relevant agencies. The approach is not only costly, but also inflexible in two senses: first, it cannot show the tourists' dynamic behaviors timely; second, data has to be re-collected if the study plan has changed. For a concrete example, the route (by public transport) to Singapore zoo recommended on the zoo's official website was not optimal, or even a bad choice for visitors staying in the west of Singapore for many years. By gathering quite a few of complaints, the website launched the interactive route recommendation module since last year. If visitor's travelling pattern could be monitored and analyzed, such services would be improved much earlier.

Despite of the existence of Hop-on Hop-off tour services, the public transport, including MRT and bus, is still the first choice by visitors to Singapore [13]. Thank to the highly efficient public

KDD'14, August 24–27, 2014, New York, NY, USA.
Copyright 2014 ACM 978-1-4503-2956-9/14/08 ...$15.00.
http://dx.doi.org/10.1145/2623330.2623352.

[1] Mass Rapid Transit (MRT) is the subway system in Singapore. The detailed description of MRT and Singapore's public transport system can be found in Section 2.

transport system, places of interest that are found widely across the Singapore island are conveniently reachable by public transport at very affordable prices. Every time when a local or a tourist rides a bus or MRT, a record that describes the time, location, fare and other information is collected in the system. Thus we see, the public transport data contain records of tourists among the records of locals. Further, if we can accurately identify these tourist records from the public transport data and analyze them, we can offer an innovative approach towards tourists' behaviors study. Compared to traditional survey-based approaches, analyzing public transport data allows complete temporal-spatial tracking of the tourists for more accurate behavior study. Furthermore, timely results can be obtained as long as the data is up-to-date.

Nevertheless, solving the tourists identification problem is non-trivial. It requires not only a comprehensive understanding about Singapore and her public transport system, but also to overcome several technical challenges: the public transport data is anonymous and the tourists records only constitute a small portion of the public transport data, yet we have to accurately differentiate them from tons of other commuter records.

1.3 Contribution and Impact

In this paper, we propose the innovative application of machine learning techniques in tourists identification from the public transport data. The public transport data that we use are provided by LTA of Singapore, and includes both MRT and bus riding records. The algorithm that we propose is based on reinforcement iterative learning: with the prior knowledge on the attractiveness of MRT stations to the tourists, which is derived from the data, and a small set of labeled data, we initiate an iterative learning process to infer potential tourists from the whole population and also update the ranking of station in each iteration. We demonstrate our model and algorithm outperforms the state-of-the-art classification methods. Furthermore, we develop an interactive web-based system that periodically discovers tourists from up-to-date public transport commuters and visualize their travel patterns to LTA and other partners from tourism industry, for better services and planning.

The research result has been recognized by LTA, and directly led the foundation of the I^{2}R-LTA Joint Laboratory, which is a long-term collaboration between I^{2}R and LTA for transport data analysis. This work also attracted interest from agencies and companies in the tourism-related industry in Singapore.

1.4 Organization

The rest of the paper is organized as follows: In Section 2, we describe the tourists identification problem that we are trying to solve. The details of our approach for the problem is presented in Section 3. We present experimental evaluation on our algorithm and other competing algorithms with real data in Section 4. Then we show a demo of the web-based system that we developed based on our learning algorithm in Section 5. We review related work in Section 6. Finally conclude this paper in Section 7.

2. PROBLEM DESCRIPTION

Before we introduce the tourist identification problem, we give an overview of Singapore's public transport system and the public transport data we own, which should be helpful for readers to understand our problem.

2.1 The Public Transport System and Dataset

In Singapore, the public transport system includes the subsystems of subway and bus. The subway system is further divided into Mass Rapid Transit (MRT) system and Light Rail Transit (LRT) system, according to the types of rails and trains. LRT uses light rails and small trains, and acts as feeder service to MRT for short-distance neighborhood railway transportation. Since the MRT and LRT share the same ticketing system, to simplify the presentation, we only use the term of MRT to represent the subway system in Singapore.

Table 1: Dataset schema

Field	Description
Card_Number_E	Card ID for this ride
Transport_Mode	BUS, LRT, or MRT
Entry_Date	Date when ride started
Entry_Time	Time when ride started
Exit_Date	Date when ride ended
Exit_Time	Time when ride ended
Payment_Mode	Method of payment
Origin_Location_ID	Starting location of the ride
Destination_Location_ID	Ending location of the ride

In Singapore, EZ-Link card is used by public transportation users to pay fare. Nearly all the residents in Singapore use EZ-Link cards tap in and tap out for bus and MRT riding. The trip fare is calculated based on the travelling distance, and deducted from a commuter's EZ-Link card when he/she tap out at a bus stop or a MRT station. For special groups of people, such as students and senior citizens, there are also concession EZ-Link cards which offers discounts. EZ-Link card is also a good choice for tourists, especially for those who stay for a few days or more and travel often, because of its convenience in use. However, buying an EZ-Link card requires a minimum payment of 12 Singapore dollars including a non-refundable 5 Singapore dollars of issuing cost. Hence, EZ-Link card might not be an economical choice for tourists who stay in Singapore for a very short period and travel with public transport in very few times. Many of such tourists may prefer using cash payment for each MRT or bus riding, as introduced later. There is also an EZ-Link day pass option for tourists, with which a tourist pays a fixed fare for unlimited rides during a whole day, and the pass turns to normal EZ-Link card after the valid period.

A commuter can also opt to use cash payment for MRT and bus riding. For MRT riding, a standard ticket need to be purchased with cash and used in a similar way as an EZ-Link card. One difference between standard ticket and EZ-Link card is that a standard ticket is disposable after one or a few times of uses. Another difference is that the per-ride price of a standard ticket is slightly higher than the EZ-Link fare for a particular trip, but the initial cost for the standard ticket is much lower compared to regular EZ-Link card as it only requires a refundable 10 cents. Therefore, the standard ticket is ideal for tourists who stay very short term or do not travel with MRT often. For bus riding, cash payment is made when a commuter boards a bus and consults the driver regarding the exact fare for his/her trip. A paper ticket will be issued to the commuter via the ticking machine after his/her payment. In Fig 2.1, we show the relationship between different payment modes and the three transport modes in Singapore's public transport system.

The public transport dataset provided by LTA consists of records generated from tap-in and tap-out activities by public transport commuters. The dataset contains records from both bus and MRT rides that are paid with regular EZ-Link card, concession EZ-Link card, standard ticket (for MRT) and paper ticket (for bus). The dataset is structured and each record follows a predefined schema. In our work, we only use a selected subset of attributes for tourists records

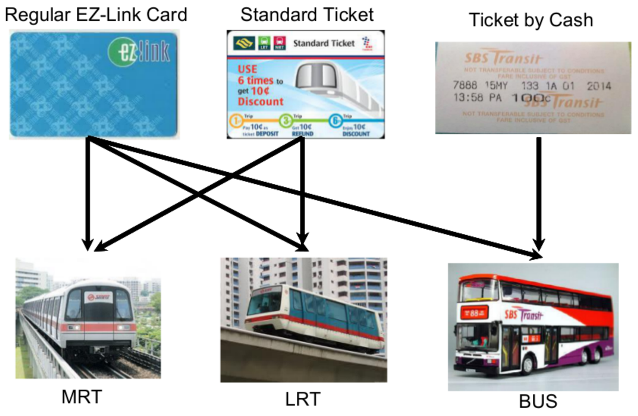


Figure 1: Singapore's public transport system

extraction, which are shown in Table 1. If the *Transport_Mode* of a record is *BUS*, the values for *Origin_Location_ID* and *Destination_Location_ID* are the corresponding IDs of bus-stops; otherwise, the values are the IDs of MRT stations. Besides of the EZ-Link records, we also have additional files for mapping a bus stop ID or MRT station ID to its names and geo-location. The *Payment_Mode* for a ride can be CSC, Pass or Standard ticket if the *Transport_Mode* is *LRT* or *MRT*. Here, *CSC* corresponds to normal adult card and the *Pass* corresponds to concession cards. As introduced above, they both belong to regular EZ-Link card. Lastly, the *Payment_Mode* for a ride can be CSC, Pass or Cash if the *Transport_Mode* is *BUS*.

2.2 Tourists Identification Problem

In this work, we target to identify records that are generated by the riding of tourists who visit Singapore from the public transport data. We assume the population can be divided into two classes, i.e., tourists and non-tourists. Sometimes, we also use locals to refer non-tourists. Tourists refer to the group of people who visit Singapore for short term, e.g., a few days, sightseeing purpose. They commonly visit places of interest of Singapore, including gardens, museums, restaurants, shopping streets, etc and stay in hotels or hostels. People who come to Singapore for other purpose such as business or medical services may also fall into the class of tourists as long as their activities satisfy our previous guideline. Non-tourists or locals are those who are not classified as tourists. Since the public transport data contains entry time, exit time, the origin and the destination, we may use these information to identify tourists from all commuters.

It might already be noticed that there are a lot of tourists records contained in the standard ticket records. This is because locals rarely use standard tickets as it is obviously not as economical, convenient and durable as EZ-Link cards. However, as a standard ticket needs to be disposed after one or a few rides, it cannot be used to reliably track the complete usages of the public transport of a tourist. On the other hand, an EZ-Link card usually stays at one's hand during one's stay in Singapore and it is good for tracking complete usages. Our focus is to identify tourist records from the entire set of EZ-Link records in the public transport data.

3. APPROACH

Our algorithm for tourist identification is a two-stages approach. In the first stage, the algorithm produces initial scores for each MRT station based on their attractiveness to the tourists. In the second stage, the scores are used to obtain initial class distributions of MRT stations. The algorithm then performs class inference for commuters, i.e., tourist or non-tourist, using an iterative process based on a graph that describes the prior class distributions of commuters and MRT stations and their interactions. There are two reasons for us to opt to use only MRT riding records rather than bus-stop ones for tourists identification. First, the standard ticket records which provide a good resource for studying tourists' travelling pattern are only available for MRT rides; Second, most places of interest in Singapore are located in walking distance. The MRT riding records should capture most of tourists' local travelling activities. Hence, it is sufficient to use them for tourist identification. Interesting, once tourists are identified, their interested places reach by bus riding can be revealed, e.g. the Singapore zoo example in Section 5.

3.1 Station Ranking

We focus on the algorithm for the initial scoring of MRT stations. The initial score s_{m_i} for the station m_i is for evaluating whether the station is more likely to be a destination for tourists or a destination for locals. Effectively, knowing someone who has visited a station with a high (or low) initial score may increase (or reduce) our belief that the person is a tourist. Thus, by considering the stations one has visited and their scores, we can differentiate tourists from locals. There are different ways to assign initial scores to stations. One ideal measurement is the probability that one is a tourist, given that he/she has visited a station. For simplicity, we denote this probability as $\Pr(t|m_i)$ in which t denotes that a commuter is a tourist and m_i denotes that the commuter has visited the station m_i. Computing the exact value of $\Pr(t|m_i)$ for each m_i is not very straightforward from the data we own. Instead, we make the transformation based on Bayes Rule:

$$\Pr(t|m_i) = \Pr(t) \cdot \frac{\Pr(m_i|t)}{\Pr(m_i)} \tag{1}$$

In the above equation, $\Pr(m_i|t)$ is the probability for one to visit m_i given that he/she is a tourist, $\Pr(m_i)$ is the prior probability for one to visit m_i and $\Pr(t)$ is the prior probability for one to be a tourist. As we see from the expression, the terms that affect the value of $\Pr(t|m_i)$ are $\Pr(m_i|t)$, $\Pr(m_i)$, and $\Pr(t)$. While $\Pr(m_i|t)$ and $\Pr(m_i)$ are dependent on the station, $\Pr(t)$ is invariant with all stations. The effect of $\Pr(t)$ can be considered as a linear scaling which does not affect the relations of the scores for different stations. In the following, we show how to estimate $\Pr(t|m_i)$ are $\Pr(m_i|t)$, $\Pr(m_i)$, and $\Pr(t)$, respectively.

The estimation for $\Pr(m_i|t)$: Ideally, to estimate $\Pr(m_i|t)$ we need to use data from tourists and summarize how often they visit each station so as to estimate the probability that they visit a particular station m_i. From prior knowledge, we already know that a lot of tourists choose to use standard tickets when taking MRT in Singapore. So it inspired us that we may obtain an estimation of $\Pr(m_i|t)$ from the standard ticket records. The main problem with using standard ticket records as tourist records is that it is incorrect to assume all standard ticket records are from tourists. Occasionally, a local may also use standard ticket in case he/she forgets to bring a regular EZ-Link card or does not have enough cash to top up the EZ-Link card that has insufficient balance. We would expect such occasions are infrequent. However, since the total rides from locals are much larger than the total rides from tourists, even if a tiny fraction of rides from locals turn to use standard tickets, they may still constitute a considerably large portion to the total standard ticket records at a station. Thus, if we need to compute $\Pr(m_i|t)$

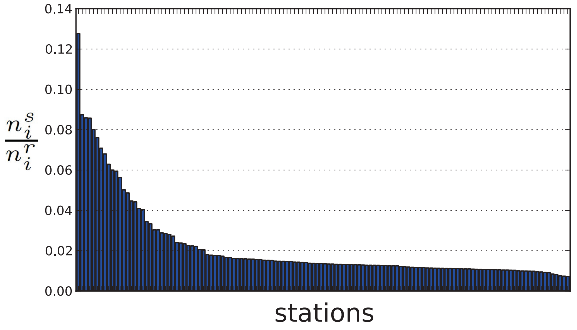


Figure 2: Stations ordered in descending c_i^s/c_i^r order

with the standard ticket records, we have to isolate the effects of locals.

To do that, we can model a parameter θ which is the expected ratio between the number of standard ticket records and the number of regular EZ-Link records from all trips made by locals. The θ is the parameter for describing how frequent a regular EZ-Link card local may turn to use standard ticket due to various reasons. For example, for a particular station m_i, we assume the total number of rides made by locals with regular EZ-Link card is 10,000. Then the number of rides from locals paid with standard tickets is estimated as $10,000 \cdot \theta$ with expected value. As we see, the θ modeling is about the behaviours of locals and is independent with stations, thus it is reasonable to assume that θ is stable for each station. Let n_i^r be the number of regular EZ-Link records and n_i^s be the number of standard ticket records at station m_i. With a fair estimation of θ, denoted as $\hat{\theta}$, we can then estimate the number of tourists who use standard tickets at m_i with its expected value $n_i^t = n_i^s - n_i^r \cdot \hat{\theta}$. In the following, we turn the attention to obtaining $\hat{\theta}$.

Table 2: Examples of non-tourist stations

Name	n_i^s	n_i^r	$\frac{n_i^s}{n_i^r}$
Marymount	6218	629435	0.009879
Yio Chu Kang	20361	2067636	0.009847
Cove	1817	189873	0.009570
Buangkok	7454	787463	0.009466
Layar	345	37211	0.00927
Oasis	489	53696	0.009107
Labrador Park	2473	292858	0.008444
Tongkang	1295	158299	0.008181
Compassvale	2705	358175	0.007552
Dover	8963	1247247	0.007186

Our idea for estimating θ is based on a key observation that tourists are very unlikely to visit stations that are located in residential areas with few shopping, hotel and restaurant facilities. Having some familiarity with the Singapore city, one can identify a few number of these stations easily. And yet, the dataset shows that there are still certain standard ticket records whose destinations are among these stations, which convinces us that these records mainly come from locals instead of tourists. Thus, by studying the ratio between the number of standard ticket records and the number of regular EZ-Link card records, we could have a fair estimation of θ. In Table 2, we show examples of these non-tourist stations with the count of standard ticket records n_i^s and the count of regular EZ-Link records n_i^r at each station and the ratio between the two. Our second observation is that, the one with the least value of $\frac{n_i^s}{n_i^r}$, which is the Dover station (the last row of the Table 2), gives the closest estimation of θ. The reason is that the smaller the $\frac{n_i^s}{n_i^r}$ value is, the less influence tourists have on the standard ticket riding records. For the information of readers who are not familiar with Singapore, Dover is a MRT station that is located next to a Polytechnic of Singapore and there is no tourist related facilities in the surrounding area. It can be imagined that most of the commuters, who exited at this station, are students or staffs from the Polytechnic. This rationale can also be explained from Figure 2, which shows the ordering of stations based on decreasing order of $\frac{n_i^s}{n_i^r}$. The values of $\frac{n_i^s}{n_i^r}$ are much higher for the stations in the first quarter than the rest stations. It indicates that the stations towards the end are non-tourist stations and their values of $\frac{n_i^s}{n_i^r}$ are getting closer the real of θ. As we see from the Table 2, the minimum value of $\frac{n_i^s}{n_i^r}$ is 0.007186. Thus, we use $\hat{\theta} = 0.007186$ as an estimation of θ, and it can be interpreted that in average 7 out of 1007 trips with MRT made by locals are payed with standard tickets. Knowing $\hat{\theta}$ allows us to estimate the number of standard tickets from tourists n_i^t at each station with the *maximum likelihood principle*, and hence estimate $\Pr(m_i|t)$ with:

$$\hat{\Pr}(m_i|t) = \frac{n_i^t}{\sum_i n_i^t} \text{ where } n_i^t = n_i^s - n_i^r \cdot \hat{\theta} \quad (2)$$

The estimation for $\Pr(m_i)$: The estimation for $\Pr(m_i)$ is much more straight-forward than $\Pr(t|m_i)$. It describes the overall probability that one may visit m_i regardless one is a local or tourist. Again the approach is based on the *maximum likelihood principle*: we count the total number of regular EZ-Link records and standard ticket records for each station m_i and divide by the total number of regular EZ-Link records and standard tickets records for all stations. Thus our estimation for $\Pr(m_i)$ is:

$$\hat{\Pr}(m_i) = \frac{n_i^s + n_i^r}{\sum_i n_i^s + n_i^r} \quad (3)$$

The estimation for $\Pr(t)$: As we have already argued, $\Pr(t)$ is a constant factor that does not affect the rankings or the relations between scores of different stations. Still, we need to estimate it for our iterative refinement algorithm to work well. Effectively, to our iterative refinement algorithm, $\Pr(t)$ acts as a tunable threshold which allows us to adjust how restrictive our algorithm is in selecting tourists records. We emphasis that a coarse estimation for $\Pr(t)$ for initial score assignment is sufficient for our algorithm to work well, as in the iterative refinement process the scores for each station will be updated to more accurate values. From our previous computation, the number of tourists arriving at an MRT station m_i using standard tickets can be estimated, i.e., n_i^t. Thus we estimate the total number of tourists arriving m_i, including both EZ-Link card users and standard ticket users, as $2n_i^t$, without any other knowledge on the percentage of the users for each type of ticket. Then our estimation for $\Pr(t)$ is:

$$\hat{\Pr}(t) = \frac{\sum_i 2n_i^t}{\sum_i n_i^s + n_i^r} \quad (4)$$

Based on the estimations of $\Pr(m_i|t)$, $\Pr(m_i)$ and $\Pr(t)$, we assign initial score for station m_i with our estimations using:

$$s_{m_i} = \hat{\Pr}(t) \cdot \frac{\hat{\Pr}(M_i|t)}{\hat{\Pr}(M_i)} \approx \Pr(t|M_i) \quad (5)$$

In Table 3 we show the top 15 stations based on the decreasing order of the initial scores. They are all famous POIs in Singapore. As we can see from the result, the top ranked station is Changi Airport which showing that visiting the airport gives the highest confidence that one is a tourist. This result perfectly matches our expectation as most tourists end their trip in the airport whereas locals do not often visit the airport. It can also be observed from the ranking that famous shopping places such as Orchard, Bugis and the CBD area City Hall are in the middle to end range of the ranking, although they are must-visit places in Singapore. Since these are also the places that are frequently visited by locals, they are not ranked so high in our scoring scheme. In other words, visiting Changi Airport gives more confidence than visiting Orchard or City Hall to infer an MRT rider as a tourist, rather than a local.

Table 3: Top ranked stations based on initial scores

Name	s_{m_i}
Changi Airport	0.213668
Marina Bay	0.145012
Clarke Quay	0.144702
Bayfront	0.128008
Little India	0.118879
Chinatown	0.113837
HarbourFront	0.106443
Bras Basah	0.104787
Esplanade	0.099637
Orchard	0.098623
Lavender	0.093104
Farrer Park	0.081844
Promenade	0.079080
Bugis	0.070973
City Hall	0.064815

3.2 Label Inference

3.2.1 Problem Representation

So far we have obtained the initial scores of MRT stations. In this section, we propose a graph structure, namely the Station-Commuter Relationship (SCR) graph, to encapsulate all the prior knowledge such as MRT stations and their initial scores, existing labeled and unlabeled commuters in the data, and their relationships. By representing our prior knowledge in the SCR graph, we also cast our problem into a node-labeling problem in the SCR graph. Specifically, in a SCR graph (Fig. 3), we define and distinguish two types of nodes for commuters: T are the nodes for the target users whose classes are unknown and we want to assign category class (tourist/non-tourist) and L are the nodes represent the commuters who already have high-quality classes, e.g. by human labeling. Usually, the number of labeled commuter nodes in L are much smaller than the number of unlabeled commuter nodes in T. For each $t_i \in T$, it maintains a class distribution, which is initialized to $[\hat{\Pr}(t), 1 - \hat{\Pr}(t)]$ (c.f. Equation 4) at the beginning for the learning algorithm. The element of the first dimension of the distribution class for t_i describes the probability for t_i to be a tourist and the element of the second dimension is the probability for it to be a non-tourist. Similarly, each $l_i \in L$ also maintains a class distribution with the same interpretation. Differently from t_i, since the class of each s_i is already known, it is initialized to $[1, 0]$ or $[0, 1]$ depending whether the commuter of s_i is a tourist or a non-tourist, respectively. Besides the nodes for commuters, there are also a set of nodes M for the stations in Singapore. Similar to the nodes for commuters, each station node $m_i \in M$ also carries a class distribution which is initially set to $\phi^0_{m_i} = [s_{m_i}, 1 - s_{m_i}]$, where the element of the first dimension represents the estimated probability for one to be a tourist if he/she visits m_i and the second is the probability for one to be a non-tourist. Let $u \in T \cup S$ be a particular commuter. The interactions between u and a station m_i, i.e. the number of times that the commuter has visited the station as a destination, represented as a weighted edge (u, m_i) where the weight w_{um_i} is the number of times that u has visited m_i.

We define and distinguish two classes. Respectively, Class 0 is the class for tourist and class 1 is the class for non-tourist. Our goal is to update the class distributions of all nodes $t_i \in T$ based on the information contained in the SCG graph, and finally assign a class label to the commuters that correspond to nodes in T.

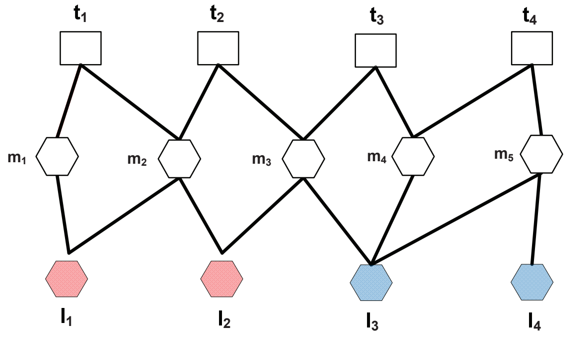


Figure 3: Example of SCR graph

3.2.2 Algorithm

The main resources for label inference in the SCR graph are the initial scores for stations, the limited number of labeled nodes in L and the interactions commuters and stations as denoted by the weighted edges. Considering the structure of the graph, we propose a label propagation approach to infer the classes for the nodes in T based on the nodes in L by taking the stations as bridges between T and L. The mechanisms we use for label propagation is based on the observation that commuters who have similar class (by the score) tend to travel to similar stations and vice versa. Specifically, we apply the following two mechanisms:

- *M1:* Update the class distributions of commuters based on the class distributions of the stations that they have visited. For example, in Fig. 4(a), t_1 is a node for an unlabeled tourist who visited HarbourFront and Changi Airport. As we see from Table 3, both HarbourFront and Changi Airport are very typical tourists places. Thus, in our label propagation, we may increase the confidence that t_1 is a tourist.

- *M2:* Update the class distributions of stations based on the class distributions of commuters who travelled to the stations. For example, in Fig. 4(a), l_1 and l_2 are tourists, and l_3 and l_4 are non-tourists. The figure shows that the station City Hall is visited by both l_2 and l_3. Since l_2 is a tourist, it may increase the confidence that City Hall is a tourist place. On the other hand, since l_3 is a non-tourist, it may increase the confidence that City Hall is a non-tourist place. In label propagation, the class distribution for City Hall is affected by the classes of both l_2 and l_3.

With the above discussion, now we use an iterative inference algorithm to solve the classification problem. The main idea behind the algorithm is that, in each iteration, we use the class distributions of stations in M to predict the labels of commuters in T. Once the labels for the nodes in T are updated, we use the nodes in T together with the nodes in S to update the class distributions of stations in M. The process ends when the number of iterations reaches a predefined threshold (In experiments, the number of iterations we choose is 150).

Algorithm 1 Iterative Inference

```
Input:
    Class label $C(l_i)$ for $l_i \in L$
    Initial scores $s_{m_i}$ for $m_i \in M$;
    Max iteration $Maxit$;
    Prior tourist probability $\hat{Pr}(t)$;
    Parameter $\alpha, \beta, \gamma$
Output:
    Class label $C(t_i)$ where $t_i \in T$
 1: // Initialization
 2: for each $l_i \in L$ do
 3:     $\phi^0_{l_i} == [1,0]$ if $C(l_i) == 0$ otherwise $\phi^0_{l_i} == [1,0]$
 4: for each $m_i \in M$ do
 5:     $\phi^0_{m_i} = [s_{m_i}, 1 - s_{m_i}]$
 6: for each $t_i \in T$ do
 7:     $\phi^0_{t_i} = [\hat{Pr}(t), 1 - \hat{Pr}(t)]$
 8: // Iterative Label Propagation
 9: for $k$=1 to $Maxit$ do
10:     // Label Propagation to $L$
11:     for each $l_i \in L$ do
12:         Update $\phi^k_{l_i}$ using equation (6)
13:     // Label Propagation to $T$
14:     for each $t_i \in T$ do
15:         Update $\phi^k_{t_i}$ using equation (7)
16:     // Label Propagation to $M$
17:     for each $m_i \in M$ do
18:         Update $\phi^k_{m_i}$ using equation (8)
19: // Output the Class Label
20: for each $t_i \in T$ do
21:     $\hat{C}(t_i) = \underset{0 \le j \le 1}{argmax} \frac{\phi^{Maxit}_{t_i}[j]}{\sum_{t \in T} \phi^{Maxit}_{t_i}[j]}$
```

The complete algorithm is shown in Algorithm 1. At steps 1-7, the class distributions for nodes in L, M, and T are all initialized, based on existing-labels, initial scoring, and prior class distribution, respectively. The iterations start from step 8. In each iteration, the class distributions for each types of nodes are updated accordingly (steps 10-18). Specifically, for nodes in L (step 11 and 12), the class distribution is updated based on both the class distributions of adjacent nodes and its class distribution in the previous iteration. A configurable parameter α is used to control the rate of update. The specific updating rule is:

$$\phi^k_{l_i} \leftarrow \alpha \cdot \phi^{k-1}_{l_i} + (1-\alpha) \cdot \frac{\sum_{m \in N(l_i)} w_{l_i m} \cdot \phi^k_m}{\sum_{m \in N(l_i)} w_{l_i m}} \quad (6)$$

In the above the edge weight $w_{l_i m}$, which is the number of times that l_i visited m, is to emphasis more important edges. $N(l_i)$ is a function for returning the neighbors of l_i in the SCR graph.

Similarity, we can define rules for updating class distributions for node $t_i \in T$, and $m_i \in M$ with different parameters β, γ for controlling the rate of update:

$$\phi^k_{t_i} \leftarrow \beta \cdot \phi^{k-1}_{t_i} + (1-\beta) \cdot \frac{\sum_{m \in N(t_i)} w_{t_i m} \cdot \phi^k_m}{\sum_{m \in N(t_i)} w_{t_i m}} \quad (7)$$

$$\phi^k_{m_i} \leftarrow \gamma \cdot \phi^{k-1}_{m_i} + (1-\gamma) \cdot \frac{\sum_{u \in N(m_i)} w_{u m_i} \cdot \phi^k_{m_i}}{\sum_{u \in N(m_i)} w_{u m_i}} \quad (8)$$

Note that larger values of α, β and γ imply greater trust in the original class distribution. In practice, the values for these parameters can be chosen as the ones that give best performance in cross-validation. with these updating rules, our algorithm effectively propagate the labels across the graph between nodes in L and T by taking the MRT stations as bridges.

Based on the class distribution, we can assign an node $t_i \in T$ to the class c at steps 20-21, such that:

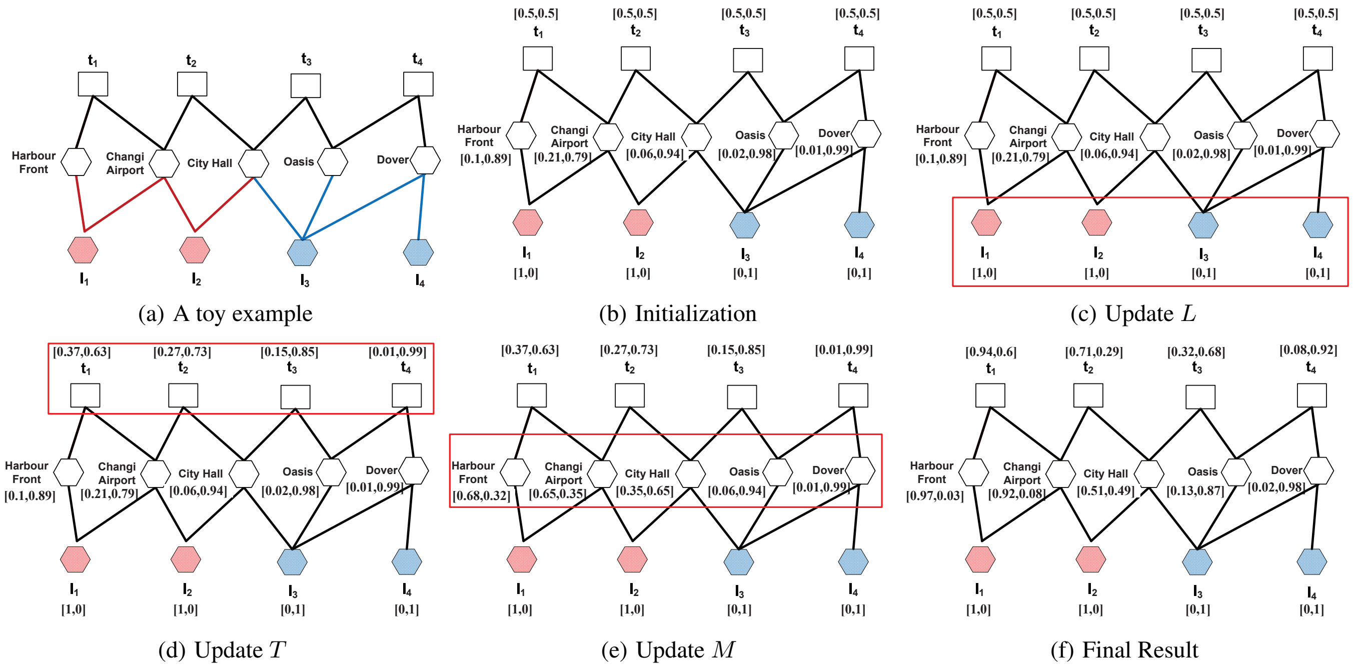
$$\hat{C} = \underset{c}{argmax} \frac{P(t_i|c)}{P(t_i)} = \underset{c}{argmax} \frac{P(c|t_i)}{P(c)}$$

Note that the final class distribution, although has been updated for describing the probability for one to be a tourist, we cannot use it for class label assignment. The reason is that the prior distribution for tourists and non-tourist are unbalanced, if we simply choose the class that gives largest probability according to the class distribution, the best strategy for a learning algorithm is to always return non-tourist. To overcome the problem, we perform normalization with the prior distributions of each class for selecting the best class.

In summary, we can benefit from this framework for the following reasons: (1) We reduce the inference problem into a label propagation problem which be handled effectively by the iterative inference algorithm on a SCR graph. (2) The assumptions adapted behind the class distribution updating rules are very intuitive which makes the framework convincing.

Back to our running example in Figure 4(a). We initialize the class distributions based on nodes types. Particularly, Nodes in L and nodes in M are assigned class distributions according to human labels and station rankings respectively. Nodes in T are assumed to have uniform class distribution [0.5, 0.5]. The initial labels are shown in Figure 4(b). In the label propagation, the class distributions for nodes in L, T and M are updated in order. Fig. 4(c)(d)(e) show the updated result for nodes in L, M and T in the first iteration. We iterate the updates 40 times for the example. The final class distributions for nodes in T are shown in Fig. 4(f). Particularly, we can calculate the category for t_2 based on $\hat{C}(t_2)$= $argmax(0.71/(0.94 + 0.71 + 0.32 + 0.08)$,$0.29/(0.6 + 0.29 + 0.68 + 0.92))$. Hence t_2 will be classified as tourist. On the other hand, t_3 will be classified as non-tourist.

The most time consuming part of the algorithm is step 9 - step 18. In the iteration, each node is visited once and the neighbors of each node are also explored. Therefore, the time complexity of the algorithm is $\mathcal{O}(k \sum_{v \in V} degree(v))$, or $\mathcal{O}(2k|E|)$ where $|E|$ is number of edges in the graph. As such, the total time complexity of algorithm is bounded by $\mathcal{O}(|E|)$. The algorithm can be further speeded up by parallel execution of class distribution updating. For example, the updating of class distributions of each node in L and each node in T based on nodes in the M can be executed concurrently, as no two nodes in $L \cap T$ are adjacent. Further, the updating of class distributions for each m_i can also be executed in concurrently as no two station nodes are adjacent.



(a) A toy example

(b) Initialization

(c) Update L

(d) Update T

(e) Update M

(f) Final Result

Figure 4: Example for label propagation

4. EXPERIMENTS

4.1 Data Preprocessing

The public transport data provided by LTA is securely stored in the A*DAX platform [2], and we are approved to use one-month records for January 2013 to test and report the performance of our learning algorithm publicly. We first do a preprocessing on the dataset. We exclude the commuters with less than 6 tuples of travelling records with EZ-Link card in the one month time, because these commuters do not have enough travelling information for our algorithm to learn and do the classification. Moreover, normally tourists with so few times of public transport riding would prefer using standard tickets rather than EZ-Link cards with extra issuing cost. Thus we do not consider the EZ-Link users with less than 6 records. Note that there may be tourists that use EZ-Link cards and travel less than 6 times, and our algorithm would overlook them. By assessing the trade-off between the noise brought in by the commuters with very few records and the unfound tourists with very limited travelling information, we tend to accept the latter. In fact, in tourists' travelling behavior study, the precision of identified tourists is more important than the completeness of tourist set by including those having limited travelling information.

After pre-processing, the data contains 1.7 million commuters with a total of 49.5 million transactions. We obtain a training set of 1 thousand tourists and 250 thousand locals, and their transactions. The tourists were manually labeled. We applied strict heuristics, e.g., the number of active days and spanning period to filter the EZ-Link card users and asked a group of people who know Singapore well to manually check each user's detailed origins and destinations, and finally labeled the tourists with high confidence. For validation purpose, we did not use these heuristics when training the model but in practice when building the actual model these heuristics can be applied as data preprocessing for initial prunning. The training data for locals were generated by special types of EZ-Link cards (e.g., student or senior) which require the holder to be present their local IC to purchase.

4.2 Experiment on Tourist Classification

We compare the proposed models with the following state-of-the-art classification methods for classification.

- SVM: Support Vector Machine (SVM) [5] is a well-known supervised classification algorithm, which uses a set of labeled data for learning.

- Fitting The Fits (FTF) [4] is a state-of-the-art iterative inference algorithm which uses both labeled data and unlabeled data. In specific, in each iteration the algorithm assigns predicted values to the observations whose responses are missing (unlabeled data), and then incorporates the predictions appropriately in the subsequent steps.

We will evaluate the effectiveness of our algorithm and the competing algorithms using standard F-measure, namely $F1$, whose formula is:

$$F1 = \frac{2 \times Precision \times Recall}{Recall + Precision}$$

The F measure can be interpreted as a weighted average of the precision and recall, where an F-measure reaches its best value at 1 and worst value at 0. In order to evaluate the average performance across multiple categories, the micro-averaging $F1$ and macro-averaging $F1$ are introduced. The micro-averaged scores tend to be dominated by the performance on common categories, while the macro-averaged scores are influenced by the performance in rare categories.

Table 4 shows the $F-measure$ of our proposed method against other methods by varying p, which represents the percentage of data used as training data. For example, when p equals 5%, it means 5% of labeled data is used to train the model, and 95% is used to validate the model. For all the algorithms, the more training data used, the better a model can be trained. It is clear that

Table 4: Results of F measure in the classification result

	SVM		FTF		I^2	
$p\%$	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1
5%	0.57984	0.8415	0.6109	0.8419	**0.6267**	**0.8504**
10%	0.5917	0.8420	0.6263	0.8464	**0.6572**	**0.8538**
15%	0.6144	0.8411	0.6441	0.8433	**0.6677**	**0.8560**
20%	0.6199	0.8480	0.6758	0.8504	**0.6962**	**0.8575**
25%	0.6286	0.8402	0.6956	0.8459	**0.7154**	**0.8549**

our method is able to achieve a higher score in both Micro and Macro F1 compared to the other two methods, for all the p values we choose. With only a limited number of labeled commuters, SVM does not perform well. The reason is that the amount of labeled data is not enough to train a good model. Actually this is the problem we aim to solve in this paper. FTF is better than SVM, since it iteratively expands training set by adding predicted labels to unlabeled data and making use of these data in the next-step learning. Our algorithm is similar to FTF, but further consider the popularity of MRT stations. This information plays as a bridge to make the label propagation more accurate.

4.3 A Case Study

We apply our classification model to the entire public transport dataset, to identify tourists and analyze some patterns of tourists. In particular, we try to extract the important rules from the identified tourists and their transactions. We apply the rule-based model C5.0 decision tree [2] onto the classification result.

Due to the space limitation, we only list several rules on the top in Table 5. Take the first rule as an example. In this rule, if a commuter's visited MRT stations satisfy the seven conditions, the commuter is classified as a tourist. By examining the rule, we can see the MRT stations of Yishun, Serongoon and Sengkang are all typical residential areas in Singapore, and far away from the city center and other POIs. It can be expected that tourists rarely visit these places, as specified in the rule. On the other hand, HarbourFront and Little India are typical POIs in Singapore. Thus the rule suggests these two places as evidences to qualify a tourist. Lastly, Raffles Place is a famous place of interest for tourists, and also the CBD of Singapore. Both tourists and locals pay a lot of visit to Raffles Place. However, most locals go to Raffles Place for work. Thus the frequency a local visiting Raffles Place is expected high. Interestingly, the rule perfectly reflects this background knowledge by specifying that if a person visits Raffles Place but less than 5 times in the month, by combining with other conditions, he/she can likely be a tourist.

5. SYSTEM DEMO

To make the best use of our analytic results, we built an interactive service for users, e.g., stakeholders, to query for their interested moving patterns of tourists. In the backend, the server maintains the most up-to-date tourist records by extracting tourist data with our algorithm when there is an update to the public transport data. The public transport data is updated periodically, and the updating frequency can be configured by system users. Initially, when a user first opens the URL of the web service, a heatmap layer[3] (Fig. 5) is loaded over the Singapore's map from Google-Map to show the popular destinations by tourists. As we see from the map, the most popular destinations, i.e., the brightest spots in the map, are located in the city center of Singapore. Other POIs such as zoo, birdpark, airport, and customs are also highlighted.

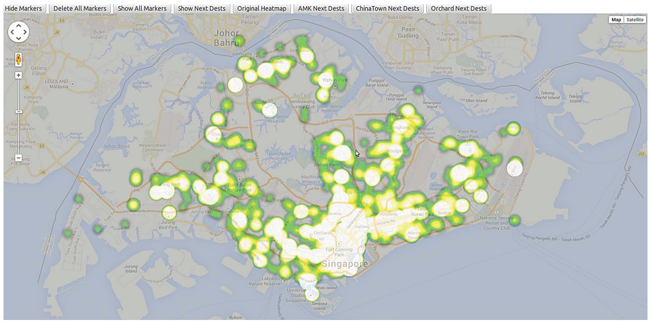


Figure 5: Heat map of tourists' favorite places in Singapore

Subsequently, a user can make queries on the next destinations of tourists by placing query landmarks with mouse clicks. The user is allowed to place multiple query landmarks at a time, and the query landmarks can be placed anywhere on the map. Once the query is issued (by clicking the "Show Next Dests" button in Fig. 5), the system automatically aligns the query landmarks to the geographically nearest bus stops or MRT stations and then return the heatmap for the next popular destinations for tourists. In addition to the heatmap, the system also returns the top 5 destination bus stops or MRT stations, by placing result landmarks, with the percentage of tourists that visit them from query landmarks. For example, in Fig. 6, we show the query results for the next destinations of tourists from the Changi airport (the red landmarks). As we see, the top destinations are all shopping, business, or leisure areas with plenty of hotels. In contrast, if we ask the same query for locals we would expect to see residential areas to come to the top. The result of this example query provide strong supports to what we find are indeed records of tourists.

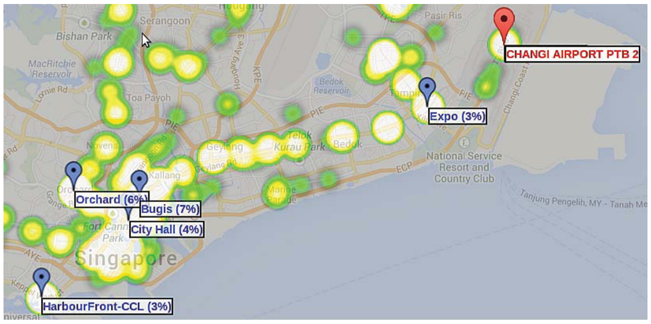


Figure 6: Destinations of tourists from the airport

[2] http://cran.r-project.org/web/packages/C50

[3] Created with gheat. url: http://code.google.com/p/gheat/

Table 5: Results of Top 5 rules in the classification result

Yishun <= 0, Serangoon <= 0, HarbourFront > 0,Little India > 0,Sengkang <= 0,Raffles Place > 0, Raffles Place <= 5 => Tourist
HarbourFront-CCL > 0, Lavender > 2, Changi Airport > 0 => Tourist
Hougang <= 0, Bishan <= 0, Orchard > 0, Kovan <= 0, Chinatown > 0, Serangoon <= 0, HarbourFront > 0, Clarke Quay > 0 => Tourist
HarbourFront <= 0, Clarke Quay <= 0, Marina Bay <= 0 and Changi Airport <= 0 => Non-Tourist
Orchard <= 0, Little India <= 0, Lavender <= 2, Marina Bay <= 0, Changi Airport <= 0 and Bayfront<= 0 => Non-Tourist

Continuing with the example, among the tourists' favorite accommodation areas, we pick Bugis, the place on the top to issue the second query, to see what places tourists prefer going to. The result is shown in Fig. 7, from which we can see the famous POIs in Singapore, such as HarbourFront (to reach the famous Sentosa island and the Universal Studio of Singapore), Orchard, Chinatown and City Hall are among the tops. Another place, Lavender also ranks very high. The reason is that Bugis itself is also a POI in Singapore, and Lavender is near to the city and there are many cheap hotels. Many tourists staying in Lavender pay visit to Bugis and come back to hotels. Thus Lavender becomes a hot destination for tourists starting trips from Bugis.

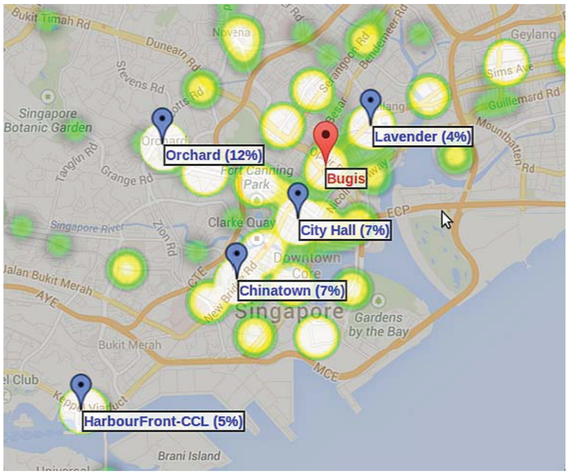


Figure 7: Destinations of tourists from Bugis

In another example, we found that the Ang Mo Kio MRT station is very frequently visited by tourists, though it is a local residential area rather than a tourism hotspot. We further investigated the reason by querying the tourist routes from Ang Mo Kio, as shown in Fig. 8. Then we found that 49% tourists went to Singapore Zoo (shown at the top left corner in Fig. 8) by bus from Ang Mo Kio. Actually, taking MRT to Ang Mo Kio and then transferring by bus was the recommended route in the Singapore Zoo's official website (this information was online for more than 10 years, and recently updated because it is not efficient for tourists staying far away from Ang Mo Kio). This explains why Ang Mo Kio became a hotspot for tourists. In fact, if this travelling pattern of tourists can be discovered earlier, the recommended routed could be revised sooner. Furthermore, if such important interchange stations as Ang Mo Kio can be discovered, services and advertisement can be set up there to target tourists.

Nevertheless, our work has opened a new door for tourists behavior analysis. There are plenty of other interesting functions one can add to the system, and the only limitation is one's imagination. In future, we would like to enhance the existing system in the following ways: 1) to support temporal constraint in the query, such as time of the day that tourists start their rides and time spent on trips 2) to support multi-hop queries to learn what places tourists may visited together in the same day.

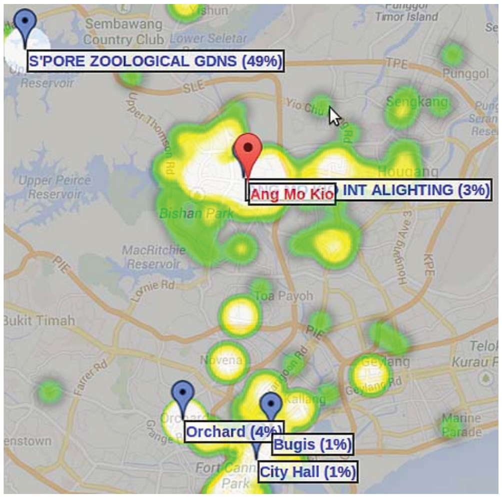


Figure 8: Destinations of tourists from Ang Mo Kio

6. RELATED WORK

6.1 Mining Public Transport Data

Because of the wide use of smart card for public transport fare collection in many countries, and the increasing amount of commuting data published for research, there are extensive works focusing on analyzing public transport data in recent years.

Studies have been carried out to analyze public transport data to improve public transport system in a city. [15] calculates the statistics from public transport smart card data that measure the transit supply and demand. This is a simple statistical collection. [7] proposes models to estimate the arrival time of bus runs based on the network transactions by 37 thousand commuters, which can be helpful for transit planning. [16] analyzes the riding records in Chicago during September 2004, and proposed evidence for service adjustment based on commuters' frequency and consistency. [6] introduces the OneBusAway tools to provide real-time arrival information for Seattle bus riders. [9] leverages on data mining techniques over the public transport data to promote personalized intelligent transport system. Other analysis on commuters' behaviors also have impact on the improvement of public transport services, as reviewed below.

Analyzing public transport commuters' riding behaviors is an interesting topic that attracts a lot of research attention from both transportation and data mining communities. [1] defines typical commuter types and analyze their trip habit and seasonal variability. [11] gets better knowledge of the behaviors of different groups of commuters according to the boarding patterns. [10] estimates the Origin-Destination (OD) matrix of bus routes in China, by applying trajectory search algorithms to track passengers' daily trip trajectories. [8] mines the commuting data and fare payment data in London, and address the problem and propose solutions to suggest best fare to commuters.

Some research works analyze the public transport data and find some result that are not directly related to the transportation system or commuters' travelling patterns. [3] detects the primary activities, as home and work activities and their locations based on the records of smart card fare payment. [14] investigates the encounter patterns of strangers on the same bus in Singapore. Our work falls in this category, however, we focus on finding foreign tourists, on which topic we did not find any existing work.

6.2 Mining Tourist Data

Data mining techniques have been used for extracting useful travel patterns from tourist-related data. Since the comprehensive tourist data is lowly available, most of existing works focus on exploiting tourist-related web data to discovery some patterns. From example, [12] tries to mine the information such as what places are often visited by tourists, how long they spend on these places, as well as the panoramic spots from geo-tagged images in Flickr. On the other hand, some researchers put effort in collecting such data. [17] recorded GPS tracks of 107 users for one year to identify the interestingness of tourist sites. Similar work also includes [18]. Compared to the existing attempts in tourist data analysis, our work uses more comprehensive dataset, i.e., the trips of all public transport commuters, and consequently faces more challenges, i.e., differentiate tourists from locals. Working on the public transport dataset, our work has high potential values in more accurately and timely figuring out real travel patterns of tourists.

There are also literatures published in the tourism research journals, which use simple analysis methods to report the behaviors of tourists and their impact to the tourism industry. Since the focus of this work is on mining transport data. We do not further review works on tourists' travelling pattern analysis.

7. CONCLUSION

In this paper, we propose methodologies to discover tourists from public transport commuters. Understanding tourists travelling behavior is important for both private and government stakeholders as it may help create new business opportunities or improve their existing services. In our approach, we first learn prior knowledge on tourists' favorite MRT stations based on the standard ticket users. With this knowledge and a limited set of labeled data, we design an iterative learning algorithm to infer tourists from the whole population of public transport commuters. We evaluate the performance of our algorithm in experimental test, and also validate the discovery result by perfectly matching the places visited by the discovered tourists with the real tourism POIs in Singapore. We further develop a web-based interactive system, based on our algorithm, for LTA and other partners from tourism industry in Singapore to use the discovery result to improve their services.

Our work has been recognized by LTA and other relevant agencies. It led the long-time collaboration between our research institute and LTA for transport data analytics. As future extension of this work, we will further analyze the travel patterns of the tourists discovered by the algorithm proposed in this paper, to meet the needs of government and industry partners.

Acknowledgement

This work was supported by the A*STAR Science and Engineering Research Council (SERC) Grant No. 1224200004 and 1224604057.

8. REFERENCES

[1] B. Agard, C. Morency, and M. Trĺępanier. Mining public transport user behaviour from smart card data. In *INCOM*, pages 17–19, 2006.

[2] N. Amudha, G. G. Chua, E. S. K. Foo, S. T. Goh, S. Guo, P. M. C. Lim, M.-T. Mak, M. C. M. Munshi, S.-K. Ng, W. S. Ng, and H. Wu. A*dax: a platform for cross-domain data linking, sharing and analytics. In *DASFAA*, pages 493–502, 2014.

[3] A. Chakirov and A. Erath. Activity identification and primary location modelling based on smart card payment data for public transport. In *IATBR*, 2012.

[4] M. Culp and G. Michailidis. An iterative algorithm for extending learners to a semisupervised setting. In *The 2007 Joint Statistical Meetings (JSM*, 2007.

[5] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, pages 1871–1874, June 2008.

[6] B. Ferris, K. Watkins, and A. Borning. Onebusaway: results from providing real-time arrival information for public transit. In *CHI*, pages 1807–1816, 2010.

[7] R. C. Ka Kee Alfred Chu. Enriching archived smart card transaction data for transit demand modeling. *Transportation Research Record: Journal of the Transportation Research Board*, (2063):63–72, 2008.

[8] N. Lathia and L. Capra. Mining mobility data to minimise travellers' spending on public transport. In *KDD*, pages 1181–1189, 2011.

[9] N. Lathia, J. Froehlich, and L. Capra. Mining public transport usage for personalised intelligent transport systems. In *ICDM*, pages 887–892, 2010.

[10] D. Li, Y. Lin, X. Zhao, H. Song, and N. Zou. Estimating a transit passenger trip origin-destination matrix using automatic fare collection system. In *DASFAA*, pages 502–513, 2011.

[11] C. Morency, M. Trepanier, and B. Agard. Analysing the variability of transit users behaviour with smart card data. In *ITSC*, pages 17–20, 2006.

[12] A. Popescu, G. Grefenstette, and P.-A. Moëllic. Mining tourist information from user-supplied collections. In *CIKM*, pages 1713–1716, 2009.

[13] Singapore Tourism Board. http://www.stb.gov.sg.

[14] L. Sun, K. W. Axhausen, D.-H. Lee, and X. Huang. Understanding metropolitan patterns of daily encounters. *PNAS*, 110, 2013.

[15] M. Trepanier, C. Morency, and B. Agard. Calculation of transit performance measures using smartcard data. *Journal of Public Transportation*, 12(1):79–96, 2009.

[16] M. Utsunomiya, J. Attanucci, and N. H. Wilson. Potential uses of transit smart card registration and transaction data to improve transit planning. *Transportation Research Record: Journal of the Transportation Research Board*, (1971):119–126, 2006.

[17] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *WWW*, pages 791–800, 2009.

[18] Y.-T. Zheng, Z.-J. Zha, and T.-S. Chua. Mining travel patterns from geotagged photos. *ACM Trans. Intell. Syst. Technol.*, 3(3):56:1–56:18, May 2012.