Identifying Activity Hubs from Ticketing Logs

Travis Povey

Rachel Cardell-Oliver
School of Computer Science and Software Engineering
University of Western Australia

Sharon Biermann
CEED Client: Planning and Transport Research Centre (PATREC)

Abstract

Public transport is a critical aspect of any modern city. Many cities are moving to using paperless smart-card ticketing systems, which provide a wealth of data about how the system is being used. This project aims to utilize SmartRider data from Perth's ticketing system, to develop an understanding of how patrons are utilizing the system. In this paper, we present a method of extracting activities from SmartRider data, identifying the likely motivation behind a patron's trip on the public transport system. We examine 8 million transactions across the Perth network during October, from which we infer the duration of a patron's stay in an area, which is used to derive the activites. Common patterns are extracted from the data, which describe typical arrival times, and durations of stays associated with activities. These activities are associated with hubs, which are focal points of the public transport network Our preliminary analysis identifies five patterns, which describe activities associated with school, work, university, shopping, and residential.

1. Introduction

Smart-card ticketing systems are becoming ubiquitous in many cities, and they offer a wealth of data that can be analysed to better understand and improve a public transport network. Perth's SmartRider system was launched in 2007, and within a year, over 70% of public transport transactions were conducted through a SmartRider (PRIA, 2012). This number has increased to approximately 77% in 2014-2015 (PTA, 2015). The popularity of smart-card systems continue to grow, among both users, and governments looking to replace paper ticketing systems.

Data collected from the SmartRider system provides accurate information about how the network is used, and can replace traditional survey methods. Currently, TransPerth uses surveys conducted in person at high volume areas of the network to assess performance and customer satsifaction (Painted Dog, 2016). SmartRider data is a largely untapped resource that could go a long way towards improving understanding of Perth's transport network.

This project focuses on understanding Perth at a regional level, by understanding travel patterns associated with activities, and how that drives traffic between hubs. An activity may be something such as working, going to school, or shopping, and part of these activities involves travelling. A hub is an area that facilitates one or more activities. In the infrastructure planning field, it is often referred to an activity centre.

1.1 Previous Research

Previous research has analysed smart-card data to understand travel patterns, and how this can be used to improve transport infrastructure. Yuan. J et al. (2014) identifies functional zones (or hubs) by analysing smart-card ticket logs, and taxi trips in Beijing, supplemented with points of interest data, which describes the number of services (such as shops, cafés, theatres etc.) in a zone. The points of interest data is used to determine the type of zone, and the transport data is used to determine which zones should be aggregated, and the relationship between zones. The current project aims to identify activities without using points of interest, and infer the activity type from its characteristics.

Poussevin et. al. (2016) focused on individual patrons of the Paris Metro network, and identified common temporal based travel patterns. The goal was to characterize each trip by an activity, which defines the reason the trip occurs. One important aspect highlighted in this paper is the seperation of the patrons into frequency bands, based on how commonly they use public transport. Frequent users are likely to travel to and from work every day, whereas someone who uses public transport only once a month is much less likely to be going to work on that trip.

2. Methodology

2.1 Defining Features

The first step is to identify features extracted from the data that are representative of an activity. We propose using a "stay" as the main feature, which is the time a patron arrives, and how long they stay within a local area. This is determined by examining trip pairs from the data; the first of which describes the arrival time and location, and the second being the departure. If the time difference between arrival and departure is 16 hours or less, and the distance between the two stops is less than 500m, it is assumed that the person undertook some activity in that area. These values were determined by examining the distribution of stays and stop pairs in the dataset.

Stays are characterized by two values; arrival to nearest hour, and duration to nearest half-hour. Each stop in the network is then described by a 56-component vector, the first 24 values representing the number of arrival times of stays originating at that stop, and the following 32 values representing the number of each duration of the stays (0 to 16, in half-hour bins). This vector represents the probability that a person arriving at that stop arrives at a particular hour, and stays for a particular duration.

2.2 Identifying Hubs

A goal of this project is to identify hubs where certain activities are undertaken. This involves grouping together stops of close proximity and similar type into hubs. In order to determine which stops should be grouped together, we first consider stop pairs. These are pairs of stops used by at least one person for a particular stay. We assume that if a person arrives at one stop, and departs from another, satisfying the conditions for a stay, then both those stops should be associated to the same hub. In order to identify highly utilized hubs strongly associated with an activity, we also require that the stop pair is utilized by a large number of patrons.

Clustering stop pairs into hubs is done spatially, grouping nearby stop pairs based on distance and density, forming coherent groups into hubs. Not every stop is sorted into a hub, only highly

utilized stops. Each hub can be associated with one or more activities that describe the reason that a person would visit it.

2.3 Determining Activities

We define the feature-vector of a hub, as the sum of all feature-vectors of all it's stop pairs. Activities will be represented as common patterns exhibited among the set of all hub's feature-vectors. In order to identify these patterns, non-negative matrix factorization (NMF) is used. This method breaks down a matrix $\boldsymbol{V}$ into a set of weights, $\boldsymbol{W}$, and basis vectors, $\boldsymbol{H}$, such that $\boldsymbol{V} = \boldsymbol{WH}$. The input to this process is the number of basis vectors to identify, and $\boldsymbol{V}$, the matrix of all hub feature-vectors. Each row in the matrix can be represented as a sum of the basis vectors, which are common patterns that appear in the matrix. This process isn't exact; the resulting matrix $\boldsymbol{WH}$ will be an approximation of the original.

The result of this process is a set of basis vectors, $\boldsymbol{H}$, which represent the activites, and a set of weights, $\boldsymbol{W}$, which represent the degree to which each hub is comprised of each activity. By normalizing the weights, we can determine the percentage mix, and dominant activity. There is no way to determine the exact number of "correct" activities to find, instead a heuristic approach is used, where the number of activities should be high enough such that the approximation is sufficiently accurate, but there should be minimal, if any, duplicate activities.

3. Results and Discussion

3.1 Extracting Stays

The current results are derived from analysing one month of travel data from October, 2016, which contains approximately 8 million transactions. Each transaction provides the following information:

- Card ID Number
- Location of Tag-On (Stop ID Number)
- Location of Tag-Off (Stop ID Number)
- Time of Tag-On
- Time of Tag-Off
- Type of transaction (standard, concession, senior)

The threshold used to determine if a pair of transactions should be used to derive a stay is based on the Tag-Off data, and the Tag-On data of the subsequent trip. A time threshold of between 1 to 16 hours was chosen. A value of 16 was chosen, as this is the typical length of an overnight residential stay. Longer durations might mean that the person travelled without using public transport, meaning we cannot know what activities are associated with that stay. 1 hour was chosen as the lower bound to eliminate transfers, as we are primarily interested in activities undertaken at a patron's final destination.

The distance threshold of 500m was chosen by calculating the 95th percentile distance between stops for all valid transaction pairs (below the time threshold, and same Card ID). This is relatively close to a commonly cited value of 400m as the maximum distance a person will walk for a bus (Daniels 2006). From the 8 million records, approximately 2 million stays are derived.

3.2 Identifying Activities

The arrival times of each stay is rounded to nearest hour, and duration to nearest half hour. In this analysis, there are only 30 half hour duration bins, as 0 – 0.5, and 0.5 – 1 were excluded. Each stop is represented by two vectors; a 24 component vector for arrival hour bins, and a 30-component vector for duration.

Each vector representing a stop is a row in the matrix used as input to the NMF process. Using the heuristic outlined in section 2.3 for determining number of activities, it was found that there are 5 distinct activities. Figure 1 shows the identified activities, separated into Arrival Hour (column 1), and Duration (column 2).

There are some recognizable patterns in these activities. Activity 2 is arrival strongly at 8am, and a duration of 6.5 – 7.5 hours. This likely corresponds to a school day activity. Activity 5 is arrival at 7 – 8am, for a duration of 8 – 10 hours, which is a typical pattern for a work day. We can also see residential stays (overnight) in activity 3, people arriving at 4-6, and staying for 13-15 hours. Activity 4 represents midday, short duration outings. This could be activities such as shopping, lunch, and general errands. Activity 1 includes mid to late morning arrivals, for a spread of durations. This activity is commonly associated with stops around universities.

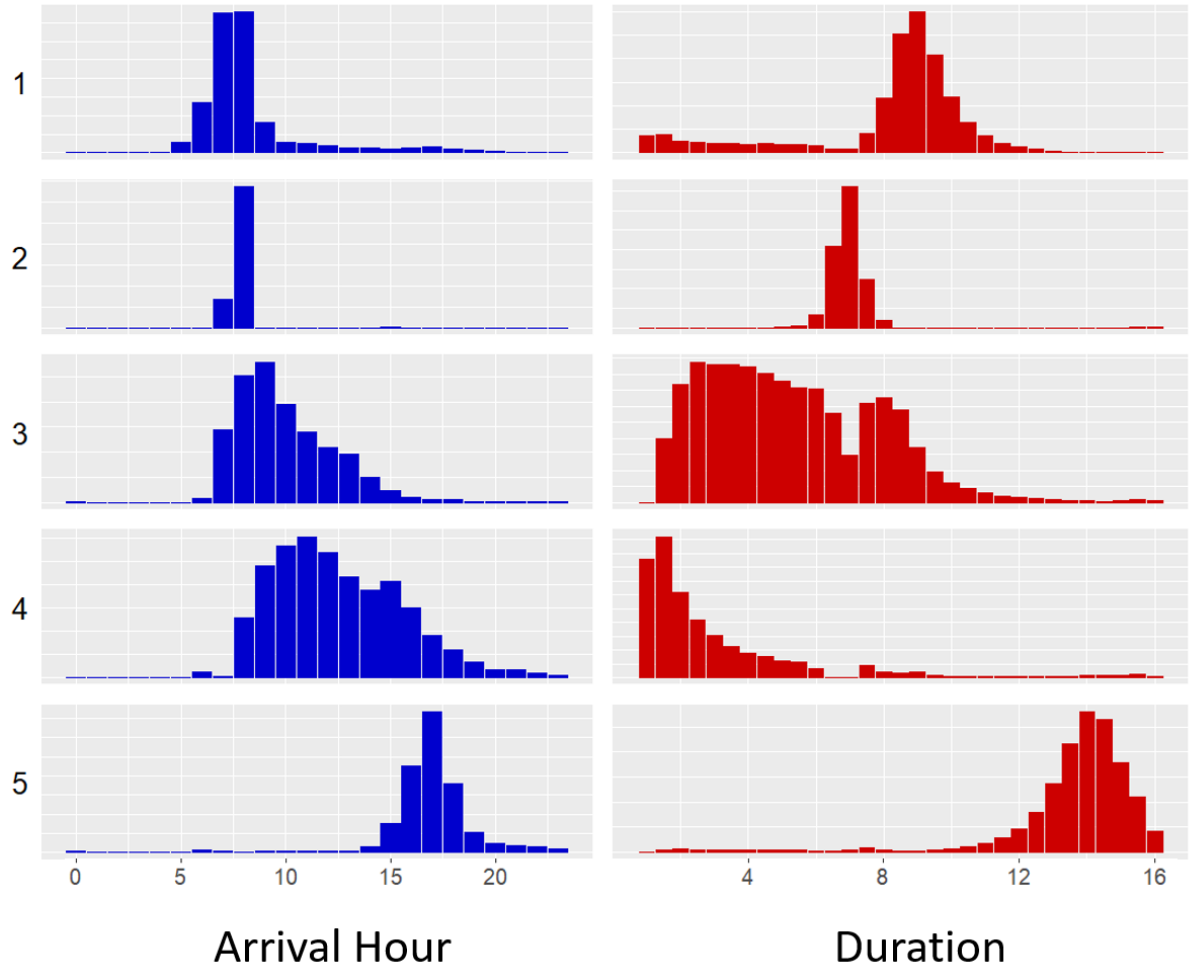

Figure 1 Activities identified from NMF on the matrix of stops. Arrival Hour is seperated into 24 one hour bins, and Duration into 30 half-hour bins (1 hour to 16 hours)

Table 1 Various hubs selected from around Perth, and the percent composition of the identified activities.

Region	Description	Volume	1	2	3	4	5
Mt Claremont	Residential Zone	1905	12%	20%	18%	9%	41%
Osborne Park	Industrial Zone	24550	53%	4%	10%	2%	31%
Karrinyup Shops	Shopping Mall	2631	13%	7%	18%	55%	7%
Perth CBD	Business District	203714	84%	4%	11%	~0%	~0%
UWA	University	72493	9%	8%	69%	11%	3%

3.3 Activities Associated with Hubs

As we are not using points of interest directly, inferring the real activity associated with each derived basis vector is mainly done using knowledge about the area around stops associated with each particular activity.

In order to gain some insight into these activities, and validate our understanding, we examine the weights associated with some identified hubs. Table 1 outlines a selection of hubs with predictable results, and their corresponding composition of activities. The results for Karrinyup Shops, Perth CBD, and UWA strongly match what would be expected of these regions.

4. Conclusions and Future Work

We have identified hubs, which are areas of that attract a large number of commuters via public transport, by analysing the SmartRider ticketing logs. These hubs have a socio-economic function, which determines the demand for transport to and from these locations.

Activities associated with hubs has also been derived from the SmartRider data. These are indentified as common usage patterns to a large number of hubs. Five distinct patterns were discovered, which we reason are associated with the following activities: work, school, shopping, university, and residential. Each hub is comprised of a mix of activities, which is representative of its function. This mix was examined for several hubs, and compared to expected results.

The next stage of this project, is to further refine the process of identifying hubs, and applying the analysis to a larger dataset. Data from different months will be analysed to test the robustness of the algorithm, and determine if the hubs are uniform or change in time.

5. Acknowledgements

I would like to thank the rest of the research group in this project, the sharing of ideas and results has helped progress greatly, Chao Sun for his advice and assistance in aquiring census data that has been invaluable to the project, and Graham Jacoby who helped me gain an understanding of transport modelling and his mentorship during my internship. Lastly, I would like to thank PATREC for supporting the project, and the Public Transport Authority and Transperth for providing the data that made this research possible.

6. References

Daniels R. & Mulley C. (2006). Explaining Walking Distance to Public Transport: the Dominance of Public Transport Supply *World Symposium on Transport and Land Use Research*

Painted Dog Research (2016). *Passenger satisfaction monitor* [online] Available at: http://www.transperth.wa.gov.au/About/Surveys-Statistics/Passenger-Surveys [Accessed 14 August 2017]

Poussevin, M. Tonnelier, E. Baskiotis, N. Guigue, V. & Gallinari, P. (2016). Mining ticketing logs for usage characterization with nonnegative matrix factorization. *Big Data Analytics in the Social and Ubiquitous Context* **9546** pp. 147–164.

Public Transport Authority (2015) *Annual report*. [online] Perth: PTA Available at: http://www.transwa.wa.gov.au/Portals/0/Repository/PDfs/PTA%20Annual%20Report_2015-16_WEB.pdf [Accessed 13 August 2017]

UTS Library (2012). *Public Relations Institute of Australia. Launch of SmartRider* [online] Available at: http://www.lib.uts.edu.au/gta/14212/launch-smartrider [Accessed 13 August 2017]

Yuan, N.J. Zheng, Y. Xie, X. Wang, Y. Zheng, K. & Xiong, H. (2014) Discovering urban functional zones using latent activity *trajectorie IEEE Transactions on Knowledge and Data Engineering* **27** pp. 712-725.