Tracing Individual Public Transport Customers from an Anonymous Transaction Database

G. Tseytin, International Business Machines, Rational Software, California
M. Hofmann, M. O'Mahony, and D. Lyons, Trinity College, Dublin, Ireland

Abstract

Data mining concepts are used frequently throughout the transportation research sector. This article examines the concept of the market basket technique as a means of gaining more insight into public transport users' demands. The article proposes a method that uses various data attributes of passenger records to infer the same customer in a different week (i.e., attempts to track the same customer from week to week). The general idea behind the measure is that if two records are considered similar, ideally every trip in one customer record should have a close counterpart in the other record. The research develops a similarity function designed to maximize the percentage of positive ticket identification over a number of weeks. Once similarity has been established, customer travel patterns can be useful in helping the operator identify new routes, new timetables, and strategic decisions in relation to satisfying public transport customer demands.

Introduction

This study is in response to the suggestion from McCarthy (2001), who argued that regular customers of a supermarket might be recognizable from patterns of their choices registered in Electronic Point of Sale (EPOS) data, and that this would help determine their long-term histories and behaviors (Chen et al. 2004). Obviously, it depends on the range of options available for each customer and on the total number of customers (e.g., in the case of a fast food restaurant offering five types of sandwiches and five types of drinks and servicing 1,000 persons daily, it may be difficult to recognize a person by his or her pattern of choices).

An attempt is made here to trace individual customers from an anonymous transaction database. The aim is to infer relations of passenger behavior that have not been noticed or at least have not been confirmed previously. Finding potential relationships among the entities that are not directly represented in the data are considered to be as important as relationships of entities that are directly represented in the data. For example, can the travel patterns of bus passengers tell us about their work routine, shopping, or spare time behavior? Mahmassani (1997) elaborates on the importance of the dynamics of commuter behavior and provides an overview, focusing on day-to-day dynamics.

The main focus of this article is to develop a method that facilitates finding record sets of routine passengers, which then can be used to further analyze passenger behavior and dynamics. The article provides a brief background of the research project and elaborates on the dataset used as an input source. A novel method that measures similarity between passenger records is then introduced. Finally, the article presents the results after applying the method to a subset of the entire data source.

Overview

In this study, magnetic strip card tickets from a public transport operator are considered. The operator provides bus services in a medium-sized European city. Train services are provided by another organization within the same group of companies. There is a predominant arterial movement of public transport services toward the city center in the morning peak periods, satisfying a well-recognized demand, and out of the city in the evenings. The tickets are issued by the public group of companies, of which the operator is one. There was no competition in the market at the time of data collection for this research, either from other bus companies or other modes such as rail.

This type of ticket is generally the primary source of passenger data (Boyle 1998). Wayfarer has manufactured the registration system used by the operator. A magnetic strip card reader at the entrance of a bus verifies a ticket; its serial number is copied into the internal memory of the device, and then onto a magnetic tape.

Other events registered on the same tape are the start of a bus journey and arrival at a specific stage (stages are selected bus stops); random stops between stages are not registered. The date and time of day, type of ticket, and route number are registered along with these events. Further, the data from every bus are copied into the transactional database.

Although there are many different types of prepaid magnetic card tickets, we will consider only weekly types (valid for a single week, starting on Sunday) and monthly types (valid for one calendar month). Within a week, a customer is not anonymous because all trip records for the same customer carry the same serial number of the ticket. All such trips taken together are comparable to a basket of items bought from a supermarket in a single visit. But, in the next week, the same customer, who is expected to use the same type of prepaid card, will have a different serial number. The question is whethercan be identified weekly ticket users from different weeks by analyzing their trip patterns.

The segmentation of the permanent stored data is on the transactional level; that is, data are stored permanently for each passenger boarding (Furth 2000). This applies regardless of whether the passenger pays with cash or has a prepaid magnetic strip card. Each piece/attribute of data is recorded as a 20-character string stored in ASCII text form. Data for a single day varies from 3 to 6 MB, depending on whether it is for a weekday, weekend, or public holiday. The file for each day averages roughly 74,000 pre-paid ticket validations.

Monthly tickets, largely similar to weekly tickets, provide an opportunity to verify whatever techniques we propose for customer identification, because they retain the same serial number throughout the month. Of course, we have to exclude weeks spanning two months, which leaves us with verification material for three (sometimes four) consecutive weeks. This article presents the results obtained for weekly portions of customer records for monthly ticket types, where the accuracy of the results can be evaluated from known customer identities. Evaluation of the same techniques for weekly types will be discussed as a separate problem.

The Data

The general course of processing is as follows. The "raw" data from a number of daily files are scanned sequentially. Dates and times of day contained in the records for the start of a bus journey and for the arrival at each stage are propagated to ticket records, along with the route number, direction number ("0" or

"1"), and stage number (unique bus stop ID). Some corrupt data can be rejected at this stage. The type of every ticket is examined, and only tickets of selected types go to further processing. The enhanced ticket records are then split by weeks, and, finally, the week files are sorted by customer numbers (the ticket type is treated as part of the customer number), whereby they can be split into weekly records of individual customers. A weekly record consists of a sequence of trips; each trip is documented by day of the week, time of day, route number, direction number, and stage number. There is no information about where the customer alighted. The average number of trips per week per passenger is approximately 13.

The next step is to convert the route and stage data to geographical coordinates so we can see for any two trips if they started at close locations or not. Geographical coordinates of each stage (there are approximately 1,000 stages) are known. From the direction of the trip (one of the two alternatives), we can derive the list of stages ahead of the boarding stage and approximate the intended direction of the customer. Some very short weekly records (three or fewer trips per week), as well as some considered corrupt (too quick a movement between geographically remote points), have been excluded. Thus, the objective of this study is to determine whether customers can be identified by their weekly "baskets," each containing about 13 trips starting from a choice of about 1,000 locations.

Figure 1 shows the contents of a sample basket from the week starting December 6, 1998, ticket type 691 (Weekly Student City zone) and ticket number 6197. The columns show the day of the week, time of departure, stage coordinates in meters, and stage name (typically, the error in the coordinates is within 50 meters, which is sufficient to enable identification of a particular bus stop).

Figure 2 shows another basket, starting December 13, with the same ticket type 691 but a different ticket number 6201. This basket was chosen to be similar to the preceding basket displayed in Figure 1, and it is a plausible hypothesis that it was the same person in both cases. However, there is no way of verifying the hypothesis because tickets of type 691 are only valid within a week. This why in the following discussion we concentrate on monthly types where the serial numbers provide a clue.

Table 1 shows the ticket types considered. The numbers given for issued tickets represents the number of customers, after the filtering, for one sample week, starting September 6.

```
1 18:10 (315103,236331) Stop A
2 10:49 (313271,238298) Stop B
2 17:48 (315103,236331) Stop A
3 10:18 (313274,238771) Stop C
3 13:45 (314876,235889) Stop D
3 14:03 (315914,234568) Stop E
3 18:05 (316801,231620) Stop F
3 18:52 (315759,235769) Stop G
4 10:31 (313272,239125) Stop H
5 10:01 (313520,238632) Stop G
5 17:55 (315103,236331) Stop A
6 10:23 (313272,239125) Stop B
6 20:21 (312888,240248) Stop I
```

Figure 1. Sample Basket of Ticket # 6197

```
1 10:27 (313274,238771) Stop B
1 18:01 (315103,236331) Stop A
2 10:34 (313274,238771) Stop C
2 18:04 (315103,236331) Stop A
3 10:31 (313274,238771) Stop C
3 20:41 (313274,238771) Stop C
3 23:11 (315899,235067) Stop J
4 10:35 (313271,238298) Stop B
4 18:08 (315103,236331) Stop B
5 10:36 (313274,238771) Stop C
6 10:32 (313272,239125) Stop C
6 11:19 (314876,235889) Stop D
6 12:17 (315914,234568) Stop E
6 17:35 (315103,236331) Stop B
```

Figure 2. Sample Basket of Ticket # 6201

Table 1. Ticket Types

Ticket Type	*Description*	*Issued Tickets*
433	Monthly Adult Short Hop Bus/Rail	622
457	Monthly Student Short Hop Bus/Rail	1216
705	Monthly Adult City zone (Airings...)	397
710	Monthly Adult Travelwide	160

Unfortunately, most popular weekly ticket types have more customers (e.g., type 671, Weekly Adult City zone, has about 9,000 customers weekly). Hence, customer identification for those types is far more difficult than in the cases with known answers.

Measuring Similarity Between Customer Records

The simplest idea for finding the same customer in a different week is to define a measure of similarity between two customer records and then to look for the best match for a specific customer record. The general idea behind the measure is that if two records are considered similar, ideally every trip in one customer record (denoted by R) should have a close counterpart in the other record (denoted by R'). The idea of identifying similarity between customers was used prior to this work in the retail sector, but this is the first time it has been used on public transport magnetic ticket data and on public transport customers. Of course, we then have to define which single trip is considered similar to which other trip; a trip being defined by the starting location, direction, and time of day (we ignore day of week for the time being) should be defined in terms of closeness of the components. If the closeness were defined as a Boolean function with only two values, we could solve a discrete task of assigning to each trip in R a close trip in R' (a sort of assignment problem). Using a fuzzy approach, each trip in R is matched with each trip in R', producing a numeric value. This value will be high for similar trips and close to 0 for differing trips. If we add together the values for all pairs, only the pairs with a good match will contribute significantly to the sum. So, the higher the sum, the better the match.

The similarity function is defined in several stages:

- Defining the weight of a trip
- Estimating the direction vector of a trip
- Comparing two trips from different customer records
- Consolidating data per starting location
- Symmetrization
- Defining scaling factors

Some of the stages were added during experiments, but no estimation was made of the effect of every single improvement, although the general impression was that each of them improved performance slightly.

Defining the Weight of a Trip

There are two reasons to ascribe different weights to trips. One is that a trip is regarded not as an independent choice of the customer, but rather as a completion of the preceding trip because the customer had to change buses. This can be decided on the basis of the time elapsed since the previous boarding and the distance between the two locations. Distances are computed just as Euclidean distances, without reference to streets of the town or barriers such as railways, rivers, and canals. The estimation of the probable speed takes account of early morning and late evening hours when speed is higher because traffic is low, as well as of special express routes with few intermediate stops (recognized by route number); the estimated ratio of "town distance" to Euclidean distance is included in the constants used.

If the distance estimated from the time interval between the two boardings and the estimated speed turns out to be less than actual (Euclidean) distance, the weight of the trip is decreased by multiplying it by the ratio of the two distances.

The other reason for weighting is frequency of stages. A rarely used stage should contribute more to the differentiation of customers than a more popular stage (e.g., a location in the center of the town). The weight factor reflecting this is taken to be proportional to the negative logarithm of the stage frequency (the intuition behind this function is that in other tasks, based on the maximum likelihood principle, logarithms of frequencies have to be added together—no other serious reason, but a function with a similar behavior has to be chosen in any case).

Estimating the Direction Vector of a Trip

The length of the vector will correspond to the degree of certainty; it can be 1, 0.25, or 0 (if there is no information about the direction). If this trip is followed by another trip taken on the same day and the trip starts from a different location, the estimation of the direction is based on the next starting location. From all stages ahead on the same route, a stage closest to the next location is sought. If it is actually closer to the next location than to the starting location of the current trip, the direction from the current location to that stage is taken with the length of the direction vector equal to 1. If there is no information about stages ahead, we take the direction to the next starting location with the vector length equal to 0.25. If there is no next trip on the same day, the best guess for direction is the farthest stage ahead. In that case, the length of the distance vector is also 0.25.

Comparing Two Trips from Different Customer Records

A similarity measure between two trips is defined as a fraction, with the denominator based on the distances between the corresponding parameters of the two trips, so the higher the distances the lower the value.

The denominator consists of 1, plus the squared distance between the starting locations, divided by an appropriate scaling factor, plus the squared difference between the starting times of day, also divided by an appropriate scaling factor. The scaling factors are discussed later in this article.

The numerator consists of the product of the weights of the two trips multiplied by 1 plus the scalar product of the two direction vectors (thus, in the worst case, when the direction vectors are opposite and both of length 1, the similarity will be reduced to 0).

Consolidating Data per Starting Locations

Typically, several trips in a customer record have the same starting location (e.g., in daily commuting from home to work). Bringing together all data for the same starting location should help estimate the role of this location in the other customer record. For a location in *R*, we add up similarity scores for all trips in *R* starting from this location and all trips in *R′*, as defined in the preceding section, to obtain a relevant figure. In this addition, some trips clearly not related to the current location will have a contribution close to 0. It is the higher values, corresponding to similar trips, that matter, and their sum is divided by the sum of all weights from *R′* (which were used as factors in each of the constituent scores). This will estimate the role of the chosen location from *R* and is denoted by *Z*. It will be compared with *W*, the total weight of the trips from *R*, starting at the location in question.

The maximum possible value for *Z* is 2*W* (attained when all scalar products are equal to 1, and locations and times coincide), but actually it should be much lower. Matching the values of *W* and *Z* is a very sensitive point in the whole procedure. If we chose just to add up the values of *Z* for all locations in *R*, it would give an unfair advantage to an *R′* record containing too many trips starting from the same location as one location in *R*. Some customer records actually have a very simple structure, just daily repetitions of almost the same trip, and in experiments such records (as *R′*) were too often selected as best matches for *R*. Hence, we should prevent high values of *Z* from making excessive contributions to the whole score. On the other hand, too low values of *Z* showing that the location is not represented in *R′* should produce a negative effect on the total measure of similarity. In such instances, a special function of *W* and *Z* was defined for the contribution of the given location to the total score (in the end the whole sum is divided by the sum of all weights in *R* to make the result less dependent on the number of trips in *R*). We will denote the result by $\alpha(R, R')$.

To define the function mentioned in the preceding paragraph, a computation was performed based on all matches between records of two neighboring weeks belonging to the same customer. This was a linear regression of Z by W, giving a sort of expectation of Z for each given value of W. Similarly, a quadratic regression was computed for the squared difference between W and the predicted "mean" value, thus providing an estimation for the variance of Z for given W. In every case the value of Z is normalized according to these estimated values (in fact, the estimation for the variance obtained from the quadratic regression can be equal or less than 0 for small values of W, and these cases are excluded from the summation). Next, the normalized value of Z is transformed by applying a function such that it is monotone and has an upper bound (in a somewhat arbitrary manner it was chosen to be $1-e^{-1.5(Z+1)}$), and then multiplied by W.

Symmetrization

The similarity function α, as defined above, is deliberately asymmetric in R and R', and the initial idea was to look for the best match for R' based on properties of R. The asymmetry is further enhanced by the choice of the scaling factors mentioned above and is also dependent on R (discussed below). However, in the final version, a symmetric similarity function was built by the following procedure. For a pair (R, R') four values are computed: $\alpha(R, R)$, $\alpha(R, R')$, $\alpha(R', R)$, and $\alpha(R', R')$. By computing such vectors for a number of pairs when it is known in each case whether they belong to the same customer, we build a quadratic discriminant function to distinguish between the two cases (no *apriori* probabilities are used, hence the constant member of this function is arbitrary, but this does not affect the search for the maximum). The value of this function, denoted by $\delta(R, R')$, serves as the symmetric measure of similarity.

Defining Scaling Factors

The scaling factors used above are needed to define which distance between two starting locations is essential and which difference between starting times is essential. The underlying idea is that customers can choose arbitrarily or, for insignificant external reasons, between several available starting points if they are almost equally remote from the actual (unknown to us) starting location or from one another.

Which distances are significant depends on the individual's habits and should be defined from this individual's behavior, though there is a default value of a "small distance." To infer the typical small distance for a specific customer, distances are analyzed between all pairs of points in the weekly record. By sorting them in the ascending order, we expect to find a gap between "small" and "big" distances (we have a preset upper bound above, so only the gaps below the limit are considered). We are looking for bigger gaps, but only if a sufficient number of distances are below the gap. We multiply the size of the gap by the total weight of distances below it, and take the lower end of the gap with the highest product (the weight of a distance is the product of the weights of the two trips from which the starting points were taken). If this procedure yields 0, the default value is used. The value obtained is the value by which the distance between the two locations is divided before squaring and adding to 1, as mentioned earlier. (A more theoretically sound alternative to this primitive approach would be a kind of cluster analysis of the set of distances, but many clustering procedures are also based on ad hoc choices.)

For time differences a similar approach is used, but a distinction is made between morning trips on weekdays, which are expected to follow a more regular pattern, and all other trips. So three time bands are defined: the division between morning trips on weekdays and other trips on weekdays, the scaling factor for time differences of general trips, and another scaling factor for morning weekday trips. In all three cases, the same approach is used as for distances, with a priori upper limits and default values and the process of finding the "best gap."

The choice of scaling factors, based on a single individual, is one more source of asymmetry between *R* and *R′*, because only *R* is used to define the values.

Finding Best Matches among Other Customers

As stated above, the simplest idea of finding the same customer in another week is, for a given customer *R*, to find the *R′* among the other week's customers that maximizes the value of $\alpha(R,R')$, or alternatively, of $\alpha(R',R)$ or $\delta(R,R')$. Table 2 presents the results for the week starting September 6, 1998.

Two obvious observations can be made from this table. First, the fewer the customers in a type, the better the results (with smaller choices it is harder to err). Second, of the three types of similarity measures, the best results are exhibited by $\alpha(R,R')$, the first function (which was designed with this type of search in mind).

Customer behavior might substantially change in the next week. The same customer may be simply absent in the other week (because he or she had less than four trips). Or, even if the same ticket is present in the other week, the pattern of usage might differ from the current week beyond recognition (for example, someone else was using the ticket while the owner was not in need of it). Figure 3 shows an example for a monthly ticket of the type 433, serial number 233, in the weeks starting September 13, 1998, and September 20, 1998.

Table 2. Results for Week of September 6, 1998

Ticket Type	*433*	*457*	*705*	*710*
Number of customers	622	1216	397	160
Same customers present in the next week	528	1007	340	147
Percentage of all customers	*84.9*	*82.8*	*85.6*	*91.9*
Best match by $\alpha(R,R')$ is correct	354	490	273	127
Percentage of all customers	*56.9*	*40.3*	*68.8*	*79.4*
Best match by $\alpha(R',R)$ is correct	254	268	205	92
Percentage of all customers	*40.8*	*22.0*	*51.6*	*57.5*
Best match by $\delta(R,R')$ is correct	345	445	262	121
Percentage of all customers	*55.5*	*36.6*	*66.0*	*75.6*

First week	Second week
0 19:57 (316058,234400) Stop L	2 11:51 (316058,234400) Stop L
2 20:10 (313520,238632) Stop G	4 12:09 (326283,226966) Stop S
2 20:54 (316058,234400) Stop L	4 15:07 (313489,233710) Stop T
4 20:33 (316698,239152) Stop M	5 19:38 (322699,249962) Stop P
6 19:59 (317906,247034) Stop N	6 22:11 (326058,227760) Stop Q
6 21:03 (316058,234400) Stop L	6 22:53 (316043,234465) Stop R

Figure 3. Example of Monthly Ticket Data

It would be interesting to estimate the number of such cases, but to compute it from the data, we need a formalization of the meaning of "differ beyond recognition." For a fair assessment of the efficiency of our method, this definition should be independent from the functions we use in the search. Surveying examples of pairs with the lowest values of the similarity function shows that very often intuitively we can observe some similarity even if the function gives a low value.

In respect to customers with weekly tickets, their usage behavior may differ from that of monthly ticket users. It was actually observed that the percentage of customer records discarded due to low usage is much lower for weekly tickets; the obvious explanation is that if a weekly ticket user does not expect to travel much in the next week, this type of ticket will not be purchased.

The analysis presented here enables public transport companies to improve their operations in a number of ways. First, it helps the company to identify the difference in travel patterns of core users. Second, the transfers between different services allows the operator to see the demand for trips involving more than one service and could potentially provide useful information on the development of new services (e.g., in the case where there is a large number of travelers starting from origin A and ending at origin C with a transfer point at B, the operator might decide to offer services starting at A and ending at C). The data can also be used to measure the possible waiting time at transfer points and to reduce this where possible.

Conclusions

This article introduced a method that facilitates finding customer ticket IDs without knowing the identity of a passenger for longer than the validity of the ticket itself. The method uses different sets of weights that are calculated to compare weekly journey patterns of individual passengers. Data attributes such as time, distance, direction, and starting location were used to calculate a weight, which then facilitates an estimate whether the two currently compared ticket IDs originate from the same customer.

The results revealed the following two main observations:

1. The fewer the customers in a ticket type, the better the results. This observation has, therefore, a negative influence for popular ticket types or for large networks with a small variety of ticket types.
2. The $\alpha(R,R')$ achieves the best results. Using $\alpha(R,R')$ on the ticket types 433, 457, 705, and 710 resulted in a positive ticket identification of 56.9 percent, 40.3 percent, 68.8 percent, and 79.4 percent, respectively. The $\delta(R,R')$ produces the second best results, which are marginally below the results calculated by $\alpha(R,R')$.

The method successfully identifies a percentage of passengers and their ticket ID of the next validity period. Considering that behavioral studies are mostly carried out on a smaller subset of passengers, the proposed method may be sufficient to chain together passengers' weekly travel records.

References

Boyle, D. K. 1998. *Passenger counting technologies and procedures*. Washington, DC: Transportation Research Board, National Research Council.

Chen Y. L., K. Tang, R-J. Shen, and Y-H Hu Y-H. 2004. Market basket analysis in a multiple store environment. *Decision Support Systems*, in press.

Furth, P. G. 2000. *Data analysis for bus planning and monitoring.* TCRP Synthesis 34. Washington, DC: Transportation Research Board, National Research Council.

Mahmassani, H. S. 1997. Dynamics of commuter behaviour: Recent research and continuing challenges. In Peter Stopher and Martin Lee-Gosselin, eds., *Understanding Travel Behaviour in an Era of Change.* Pergamon Press, pp. 279–313.

McCarthy, J. 2001. Phenomenal data mining: From data to phenomena. Computer Science Department, Stanford University. Available at URL:http://www-formal.stanford.edu/jmc/data-mining.html.

About the Authors

Gregory Tseytin (*tseytin@acm.org*) was head of Intelligent Systems Laboratory at the Acad.V.I. Smirnov Research Institute for Mathematics and Mechanics at the University of St. Petersburg, Russia. He is currently an advisory software engineer at International Business Machines, Rational Software, California. His interests include object oriented logic and alternative formalizations of common thinking.

Markus Hofmann (*mhofmann@tcd.ie*) is a Ph.D. candidate at the Centre for Transport Research, Trinity College, Dublin (TCD), Ireland. He has focused on the utilization of transport data using data mining algorithms for several years. He has published in a number of journal and conference reports and has taught widely in the area of information technology. His other research interests include performance and level of service measures of public transport operators, data mining for transport planning, and knowledge management.

Margaret O'Mahony (*margaret.omahony@tcd.ie*) holds the Chair of Civil Engineering (1842) and is director of the Centre for Transport Research at Trinity College Dublin. She coordinates several internationally and nationally funded research projects and is a reviewer for the top transport research journals. The author of more than 90 publications, her interests include transport modelling, public transport policy, data analysis, innovative technologies applied to transport analysis, demand management, and environmental impacts of transport.

Donal Lyons (*donal.lyons@tcd.ie*) graduated from University College Dublin in 1967 with a distinction in physics and with a B.Sc in mathematics. He received an M.Sc in statistics and operations research from TCD. He subsequently worked in the Irish Dairy Board as O.R. analyst, product executive, planning executive, and systems development manager. As a lecturer in the Statistics Department at TCD, his main research interest was data mining, primarily in the area of identification of semianonymous transactions. More recently, he has been the data warehouse manager at TCD.