Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 22 (2017) 25–34

Transportation Research Procedia

www.elsevier.com/locate/procedia

19th EURO Working Group on Transportation Meeting, EWGT2016, 5-7 September 2016, Istanbul, Turkey

Definition of user groups applying Ward's method

Domokos Eszergár-Kiss*, Bálint Caesar

Budapest University of Technology and Economics, Muegyetem rkp. 3., Budapest 1111, Hungary

Abstract

Requirement of passengers regarding their journeys and journey planners were the topic of this paper. First a framework of aspects for multimodal journey planners was created, where passengers were divided into 5 user groups. In order to receive real answers a survey was conducted. Considering the main aspects (route planning, booking and payment, handled data, complementary information, supplementary information) the established user groups showed no significant differences. The connection of single aspects was investigated using correlation analysis. Therefore the need of the creation of new user groups has been arisen. Ward method, as a hierarchical clustering method, was used to create groups, where the variance within the groups is minimized. The clustering algorithm was implemented in MatLab environment working with the original survey answers. As a result 5 new user groups were presented with special features, as alternative journey planning group, visualization on the map group, dynamic data group, no mobile payment interested group, no WiFi interested group. These groups showed significant difference regarding the main aspects. Using the new group allocation, passengers were categorized according to real requirement into homogeneous user groups. In order to apply the results of the Ward user groups, the evaluation of multimodal journey planners was also performed, where requirements of the user groups were taken into account during the process

Peer-review under responsibility of the Scientific Committee of EWGT2016.

Keywords: user groups; classification; Ward method; journey planner; evaluation

1. Introduction

The topic covers a current research problem of definition of user groups in transportation. The importance of different stakeholder involvement in decision-making was highlighted by Bulckaen et al (2016). In our analysis user groups are applied to classify passengers, where user groups are called clusters and answers of users are aspects.

* Corresponding author: Domokos Esztergár-Kiss. Tel.: +36-1-463-1029
E-mail address: esztergar@kku.bme.hu

2214-241X
Peer-review under responsibility of the Scientific Committee of EWGT2016.
10.1016/j.trpro.2017.03.004

Classification enhances user experience by taking into account personal parameters and enabling targeted development of journey planners. The creation of user groups is based on answers of the users of a questionnaire, which allows knowing user requirements in a more accurate way, as it was assessed by Duleba et al. (2012). Identifying patterns in answers of users can be performed by clustering methods. Hierarchical clustering is one of the most used methods, where a set of answers is present without any pre-classified grouping (Steinbach et al. (2003)).

Clustering algorithms are used in many different fields, but also in transportation. In the paper of Chiang et al. (2003) airline passenger behavior was investigated from the economical viewpoint. Market clusters were established in order to enable companies to focus their marketing strategy on similar user groups. An analysis of service quality in transportation is conducted by de Ona and de Ona (2015), who defined qualitative aspects of public transportation services and created user groups based on age and usage habits. A CART method was used, which includes classification and regression trees. The analysis was focused on revealing the most important public transportation services for the established user groups. Becker and Albers (2015) considered indicators for public transport service quality, as convenience and cleanliness of stations, seats and safety of vehicles, ticket sale opportunities. Furthermore travel information and conditions (e.g. schedule, punctuality, number of transfers) are also taken into account. The users are grouped based on their age and trip purpose (work, leisure, shopping, services). The goal of their study is to find out the reliability of the given answers by the users. These clustering applications are interesting researches, but do not deal with travel information and user preferences. The aim of this paper is to present a solution for user group definition using a hierarchical clustering method.

2. Classification method

Correlation analysis is a statistical method to define connections between single aspects, as defined by Cramer (1946). In our research it is applied to determine the consistency and similarity of user groups regarding single aspect pairs. The means of answer values for single aspects are compared. The correlation coefficient is the proportion of the covariance and product of standard deviations of single aspects (s_x, s_y). Correlation can have values from -1 to +1, where -1 means negative (inverse) correlation and +1 means positive correlation. Values close to 0 represent independency between the single aspects.

$$corr(x,y) = \frac{cov(x,y)}{s_x * s_y} \quad (1)$$

If not only two single aspects need to be compared, other methods should be applied. Hierarchical clustering methods put the elements into clusters based on similarities among the clusters (Kaufmann and Rousseeuw (2008)). The most well-known methods are single-linkage, complete-linkage and average linkage. A more complex, but accurate method is the Ward method (Ward (1963)). This method was used for classification of users and creation of new user groups. By applying Ward method (Everitt et al. (2011)) the aim was to join elements into clusters so that the variance within clusters is minimized.

Hierarchical clustering methods can be agglomerative or divisive (Aggarwal and Reddy (2014)). Agglomerative methods collect elements into groups (clusters), while divisive methods separate the elements successively into groups. Ward method is agglomerative, thus it partitions elements into a dedicated number of clusters in several steps. First each element is independent, and then step by step more elements will be ordered to a cluster. At each step the method includes those elements, which are the “closest” (according to a metric) to the existing clusters. The number of steps may reach from 1 to n (number of analyzed elements). In case of 1 only one single cluster contains all elements, while in the case of n all elements form an own cluster. Once a cluster is created as a result of a step, the elements of the new cluster cannot be separated again. The algorithm tries to find the optimal number of clustering steps. Ward method is conservative, monotone and creates about same big groups, but is sensitive to outliers (Almeida et al. (2007)). Comparing Ward method to other methods offers higher accuracy concerning the results and minimizes the variance between elements. Hands and Everitt et al. (1987) pointed out that between five clustering techniques comparing their capability to form the original clustering structure, Ward’s method did better overall than other hierarchical methods. Blashfield et al. (1976) simulated on many datasets, that Ward’s method performed significantly better than other clustering procedures.