Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 1855–1862

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

Estimation of customer questionnaire responses from purchase transaction data using canonical correlation analysis

Natsuki Sano*, Fuminori Kimura

Onomichi City University, 1600-2 Hisayamada-cho, Onomichi 722-8506, Japan

Abstract

Customer lifestyle values measured by their responses to a questionnaire provide useful information on customer behavior that supplements customer transaction data. However, although customer transaction data are accumulated automatically, it is difficult to gather questionnaire responses from many customers. In this study, we associated customer transaction data with questionnaire response data by using canonical correlation analysis. We propose two methods to estimate the questionnaire responses of customers whose questionnaire response values are in fact unknown. The first estimates the questionnaire responses by using the mean of neighborhoods defined by canonical variates. The second method estimates questionnaire responses by solving a linear equation. We validated the proposed methods using test data where the questionnaire responses are known. The estimated questionnaire responses can be used in various applications, such as the construction of a customer behavior model and recommendation methods.

Peer-review under responsibility of KES International

Keywords: Marketing; Customer transaction data; Questionnaire; Multimodality; Canonical correlation analysis

1. Introduction

To examine customers ' purchase behavior, in many studies, e.g., that of Guadagni et al.[1], customer transaction data, such as point of sales (POS) data, have been utilized. In addition, many methods to measure customers ' lifestyle values, e.g., VALS[2], have been proposed by social psychologists. It can be expected that useful knowledge can be discovered by associating customer transaction data with questionnaire response data. However, it is difficult to administer questionnaires to all customers to collect data on their customer lifestyle values corresponding to their customer transaction data.

According to the theory of multimodality, semantic analyses of several modes can be conducted (media), such as textual, aural, linguistic, spatial, and visual resources[3]. In an early study on multimodality, Kurita et al.[4] gathered images, assigned adjectives to each image, and associated these images with adjectives by using canonical correlation analysis (CCA)[5]. They constructed a retrieval system based on an image database by using adjective words.

* Corresponding author. Tel: +81-848-22-8311; fax: +81-848-22-5460
E-mail address: sano@onomichi-u.ac.jp

1877-0509
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.230

The relation between customer behavior and customer values can be considered a type of multimodality relation. The purpose of this study was to extract the relation between the purchase behavior and values of customers and by using CCA to estimate questionnaire responses of customers who in fact did not respond to the questionnaire. The data used were real data of a clothing company. The paired customer transaction and questionnaire data were divided into training and test data. To validate the proposed methods, we estimated the questionnaire data of the test data based on the relation between purchase behavior and customer values obtained from training data and calculated the root mean square error (RMSE) to evaluate the prediction accuracy.

2. Data summary

We now summarize the apparel retail data, which were composed of transaction data and questionnaire response data. The number of transaction data records was 1,009,609, the number of customers was 101,491, and the number of product items was 421,240. The number of product item categories was 226. We aggregated the purchase frequency for the product item category. Table 1 shows the top 10 product categories by the purchase frequency value.

Item category	No. of purchases
T-shirt, cut and sewn	128,285
Pants	71,266
Knit/sweater	69,703
Shirt/blouse	61,932
One-piece dress	50,288
Sneakers	37,006
Skirt	32,500
Cardigan	28,474
Socks	23,468
Denim pants	20,638

Table 1. Top 10 item categories

With respect to the questionnaire data, the total number of customers who responded to the questionnaire was 3,144, of whom 3,117 purchased an item. The number of questionnaire items was 107, of which 5 were free description items. Thus, the customer transaction data and questionnaire response data of 3,117 customers and 102 numeric questionnaire items were used in this study. The questionnaire format is shown in Table 2.

3. Association of purchase behavior with values of customers by canonical correlation analysis

3.1. Canonical correlation analysis

We associated the category of the purchased product with the response to the questionnaire item by using CCA[5] in order to understand the relation between customers ' purchase behavior and their values. First, we summarize CCA briefly. CCA computes new variates x_i, y_i, called first canonical variates, by linear combination of two groups $q_{i,j}, f_{i,j}$:

$$\begin{aligned} x_i &= a_1 f_{i,1} + a_2 f_{i,2} + \ldots + a_j f_{i,j} + \ldots + a_l f_{i,l} \\ y_i &= b_1 q_{i,1} + b_2 q_{i,2} + \ldots + b_j q_{i,j} + \ldots + b_m q_{i,m} \end{aligned} \tag{1}$$

where $f_{i,j}$ and $q_{i,j}$ denote the i-th customer's purchase frequency of the j-th product category and his/her response to the j-th questionnaire item, respectively. l and m denote the number of the category and the questionnaire item, respectively.