Proposal of a Purchase Behavior Analysis Model on an Electronic Commerce Site Using Questionnaire Data

Ryotaro Shimizu^{*1}, Teppei Sakamoto¹, Yamashita Haruka², Goto Masayuki¹

1. Waseda University : 51-15-03 3-4-1 Okubo Shinjuku Tokyo Japan 169-8555

2. Sophia University : 3-233 Yonban-cho Chiyodaku Tokyo Japan 102-0081

**contact author's e-mail address: shi3mizu8-r@fuji.waseda.jp*

Abstract:

Due to the recent development of electronic commerce (EC) technology, sale amounts on EC sites have increased rapidly. Large amounts of consumer purchase history data can now easily be obtained, and many data-based approaches to identifying consumer purchase trends have been studied. On the other hand, since diversification of consumer consciousness such as their values and lifestyles are important for marketing research, many approaches to identifying the relationship between “consumers’ values and consciousness” and “consumers’ preferences underlying their purchase behavior” have been discussed in the existing literature. To investigate consumers’ values and consciousness, a questionnaire survey is an efficient tool. It is, however, difficult to collect questionnaire responses from all consumers due to cost issues. Therefore, it is difficult to know the values and consciousness of consumers who have not responded to the questionnaire.

In this study, we propose a new model to identify both consumer purchase behavior and their consciousness based on a latent class model estimated by questionnaire responses and the purchase history data for all users. In addition, we apply the proposed model to actual data to analyze its effectiveness.

Keywords

Probabilistic latent semantic analysis, Purchase history data, Questionnaire survey data, Purchase behavior analysis model

1. Introduction

In recent years, purchases through electronic commerce (EC) sites have become popular among general consumers in many countries, and storing large amounts of consumer purchase history data has become relatively easy for corporations (Namatame et al. 2013). Thus, it has also become common to analyze purchase history data accumulated on EC site databases, and use the information as a marketing tool (Abe and Kondo 2005; Sagawa and Hirooka 2003; Ishigaki 2010). For example, understanding the purchase trend for each customer helps to identify other items which are likely to be of interest. This kind of information can be used to make product recommendations, and to issue discount coupons that are customized for each consumer. Moreover, by clustering consumers by purchase trends and customizing customer services for each customer cluster, it becomes possible to offer services that will likely yield high customer satisfaction (Shinnou 2007).

However, although it is possible to clarify the purchase trend for each consumer solely by analyzing his/her purchase history data, it is difficult to fully grasp their values and consciousness (hereinafter, consumer consciousness) that underlies their purchase behaviors. The purchase trend of a consumer is regarded as reflecting their individual lifestyle and values. Thus, if we can analyze these relationship, that information can be utilized to develop new marketing strategies. The most common way of obtaining consumer consciousness, is through a questionnaire survey. By analyzing both the purchase history and the questionnaire data, we can identify the relationship between purchase trends and consumer consciousness and understand consumer preference as a purchasing pattern that is based on consumer consciousness. Thus, this type of analysis can be utilized for marketing.

Ishigaki et al. (2011) proposed a model that simultaneously analyzes purchase history data and questionnaire data based on a latent class model called the Probabilistic Latent Semantic Analysis (pLSA, Hoffman 1999) and divides consumers into clusters. However, a limitation of the model is that only consumers who have both a purchase history and questionnaire data can be analyzed. Thus, it may be difficult to utilize the purchase history data of most consumers since it is not possible, cost-wise, to collect questionnaire responses from all consumers who make purchases through a EC site. In fact, in the case of the this study, while the number of consumers who had purchase history data was about 100,000, the number of consumers who responded to the questionnaire survey (hereinafter referred to as “respondents”) was only about 3,000. Thus, when applying Ishigaki's model, it is impossible to analyze most consumers (hereinafter referred to as “non-respondents”) since they did not respond to the questionnaire.

This study proposes a new purchase behavior model, which analyzes both purchase history data obtained from all consumers (both respondents and non-respondents), and a limited number of questionnaire data only from the respondents. We divide both respondents and non-respondents into clusters from the perspective of the relationship between consumer consciousness and purchase history. In our model, it is possible to complement the unobserved questionnaire data (which would have been obtained if non-respondents had responded to the questionnaire) by using purchase history data and to reflect consumer consciousness in the clustering results. By conducting consumer clustering by using the proposed model, consumer groups with different purchase behaviors can be characterized by consumer consciousness, which can be used to develop various marketing measures such as product recommendations based on the characteristics of the consumer clusters. Furthermore, we show the effectiveness of our proposal by applying the proposed model to purchase history data accumulated in an apparel EC site A and questionnaire data that site A has collected from a subset of its customers.

Note: This study includes sections from an un-reviewed report (Shimizu et al., 2017) in Japanese submitted to the journal of “Communication of JIMA (Japan Industrial Management Association).”

2. Background

2.1 Probabilistic Latent Semantic Analysis (pLSA)

In this study, we propose a new model based on pLSA, which was proposed by Hoffman (1999) that represents consumers' purchase behaviors as well as consumer consciousness of non-respondents. The pLSA is a latent class model widely used for document classification (Ueda 2004; Xue et al., 2008) and recommendation systems (Ishigaki 2011). It is a probabilistic latent class model that assumes unobserved latent classes underlying the purchase behavior of consumers. In this model, the purchase behavior is represented by an event that a consumer and a purchase co-occur under a latent class, and the co-occurrence relation is expressed by a conditional probability. This model assumes a latent class (which cannot be observed) between customers and products. Customer purchase behavior is expected to reflect potential preferences for the product, and are useful to model their purchase trends. By assuming latent classes between the consumers and products, it becomes possible to construct the model representing heterogeneity of consumer preferences and product characteristics.

Here, we define the latent class set with size K as $Z = \{z_k: 1 \leq k \leq K\}$, the consumer set with size I as $U = \{u_i: 1 \leq i \leq I\}$, and the product set with size J as $X = \{x_j: 1 \leq j \leq J\}$. The graphical model of pLSA is shown in Figure 1.

Furthermore, the probabilistic model of the complete data in pLSA is expressed by equation (1).

$$P(u_i, x_j, z_k) = P(z_k)P(u_i|z_k)P(x_j|z_k) \quad (1)$$

In the learning phase, parameters $P(u_i|z_k)$, $P(x_j|z_k)$, and $P(z_k)$ are estimated to obtain an appropriate model from the training data set. However, the analytical way of estimation is not enabled (Hoffman 1999). Since the model contains latent variables z_k , which are unobserved, estimation based on expectation-maximization (EM) algorithms (Dempster et al., 1977; Goto et al. 2014) is applied. The EM algorithm in pLSA can obtain a local optimal solution by updating parameters sequentially by repeating the E-step, which is an operation for taking the expectation value of the log-likelihood, and the M-step, which is an operation for calculating parameters to optimize the expected value.

Figure. 1 Graphical model of the pLSA

2.2 Ishigaki's model

The model proposed by Ishigaki et al. (2011) is a purchase behavior analysis model using both purchase history data and questionnaire data. This is a latent class model based on soft clustering, which targets only the respondents. This model assumes latent classes for each consumer and product. Here, the latent class of a consumer is denoted by $v_s (s = 1, \dots, S)$, and the latent class of a product is denoted by $w_t (t = 1, \dots, T)$. The probabilistic model of the complete data is expressed by equation (2).

$$P(u_i, x_j, v_s, w_t) = P(v_s)P(u_i|v_s) P(w_t|v_s)P(x_j|w_t) \quad (2)$$

In addition, from the questionnaire data, the consumer consciousness score is calculated. Then, parameters $P(v_s)$ and $P(u_i|v_s)$ representing the characteristics of consumers are learned by the EM algorithm using the score as a constraint condition.

This enables the analysis of consumer purchase behavior considering consumer consciousness, which cannot be extracted from purchase history data. However, a strict requirement of the model proposed by Ishigaki et al. (2011) is that the questionnaire data be obtained for all consumers.

3. Proposed analysis method

In this study, we propose a latent class model that allows for the analysis of consumer preferences by clustering consumers considering both purchase trend and consumer consciousness from the purchase history data of all consumers (i.e., respondents and non-respondents) and questionnaire data from only the respondents. The procedure of the proposed model consists of the following two steps, which will be described in detail in the following subsections.

3.1 Understanding consumer consciousness

First, to analyze consumer consciousness, factor analysis (Nagata et al. 2001; Mizuno 1996; Okamoto 1970) is applied to the questionnaire data from the respondents. By applying the factor analysis, the multivariate questionnaire data can be synthesized into several factors, and a factor score that shows the degree of correlation between each consumer and factor can be derived. Here, let F be the number of factors, and let $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iF})^T$ be the factor score vector for each consumer u_i .

Furthermore, soft clustering based on the Gaussian mixture model (i.e., GMM) (Rasmussen 1999; Hirai 1970) is performed for the respondents by using factor scores obtained from the questionnaire data of the respondents by factor analysis. The occurrence probability of the latent class z_k obtained from the GMM is represented as $P_{\text{GMM}}(z_k)$, and the conditional probability that measures the likelihood of factor score $\mathbf{y}_i$ of consumer u_i belonging to class z_k is represented as $P_{\text{GMM}}(z_k|\mathbf{y}_i)$.

3.2 Understanding consumer preference using the proposed model

This study proposes a new clustering method considering both purchase trends obtained from the purchase history data of all consumers and customer awareness data obtained from a small number of questionnaire respondents. Specifically, we propose a learning method that reflects consumer consciousness represented as a factor score calculated by factor analysis of questionnaire data when estimating the parameters of pLSA for the clustering using purchase history data.

The proposed model is based on equation (1), and we estimate the parameters $P(z_k)$, $P(u_i|z_k)$, and $P(x_j|z_k)$ in the learning phase of the model. However, since the latent class z_k cannot be observed in the model, it is estimated by using the EM algorithm (Miyagawa 1987; Dempster et al., 1977). Thus, when considering all consumers, it is possible to examine the type of class that expresses the consumer consciousness, and the share of that particular class out of all the classes. Moreover, when estimating parameters by using the EM algorithm, consumer consciousness learned by the GMM is taken into account in the model. Specifically, we set the initial value of $P(z_k)$ as the probability $P_{GMM}(z_k)$, which represents how often each latent class appears in the GMM for factor scores. Furthermore, in the calculation of $P(u_i, x_j, z_k)$ in the E-step for respondents, we use $P_{GMM}(z_k|y_i)$ obtained from the questionnaire data. Then, $P(u_i|z_k)$ in equation (1) is assumed by equation (3). Here, e_i is defined as a variable that takes 1 if consumer u_i is a respondent, and 0 otherwise.

$$P(u_i|z_k) = \left(\frac{P(u_i)P_{GMM}(z_k|y_i)}{P(z_k)} \right)^{e_i} P(u_i|z_k)^{1-e_i} \quad (3)$$

Therefore, when estimating the assignment probability of the respondent to each latent class, it is possible to consider the consumer consciousness obtained from the questionnaire data together with the purchase trend obtained from the purchase history data. Moreover, with regard to the consumer consciousness of non-respondents, we are able to estimate the parameters by complementing the missing observations (i.e., the questionnaire data which would have been observed for the non-respondents) by using the consumer consciousness data obtained from the respondents and the purchase history data of the respondents and non-respondents.

Here, the n -th consumer and product in all N pieces of purchase history data are denoted by a_n and b_n . The details of the EM algorithm in the proposed model are provided below.

i) Initial value setting

We use $P_{GMM}(z_k)$ obtained from the GMM as the initial value of $P(z_k)$, and set the initial parameters $P(u_i|z_k)$ and $P(x_j|z_k)$ randomly.

ii) E-step

Operations for taking the expectation value of the log-likelihood are as follows.

$$P(u_i|x_j, z_k) = \left(\frac{P(u_i, x_j, z_k)}{\sum_{k=1}^K P(u_i, x_j, z_k)} \right) \quad (4)$$

$$P(u_i, x_j, z_k) = P(z_k) \left(\frac{P(u_i)P_{GMM}(z_k|y_i)}{P(z_k)} \right)^{e_i} P(u_i|z_k)^{1-e_i} P(x_j|z_k) \quad (5)$$

iii) M-step

Operations for calculating parameters to optimize the expected value are as follows.

$$P(z_k) = \frac{1}{N} \sum_{n=1}^N P(z_k|a_n, b_n) \quad (6)$$

$$P(u_i|z_k) = \frac{1}{NP(z_k)} \sum_{n=1}^N \delta(a_n, u_i) P(z_k|a_n, b_n) \quad (7)$$

$$P(x_j|z_k) = \frac{1}{NP(z_k)} \sum_{n=1}^N \delta(b_n, x_j) P(z_k|a_n, b_n) \quad (8)$$

Here, $\delta(\alpha, \beta)$ is the indicator function that takes 1 if $\alpha = \beta$, and 0 otherwise.

iv) Repetition of the E-step and M-step

Until the following log likelihood L converges, steps ii and iii are repeated.

$$L = \sum_{k=1}^K \sum_{n=1}^N \log P(a_n, b_n, z_k) \quad (9)$$

4. Evaluation experiment

To confirm the effectiveness of the proposed method, we perform an evaluation experiment of the proposed model by using actual purchase history data and questionnaire data of an apparel EC site that was provided for a 2016 data analysis competition sponsored by the Joint Association Study Group of Management Science in Japan (Data analysis competition 2016).

4.1 Experimental conditions

The observation period of the purchase history data is from April 1, 2015 to March 31, 2016, and the total number of consumers is $I = 101,501$ (of which 3,118 were respondents), the total number of products is $J = 421,290$, and the total number of purchases is 1,001,901. In addition, the questionnaire contains 107 questions. In this case study, we use only the 31 questions related to “views on fashion.” In addition, the number of factors contracting the questionnaire data by the factor analysis is set at $F = 12$, and the number of latent classes when soft-clustering respondents by the GMM is set at $K = 5$.

4.2 Experiment results

To show the effectiveness of the proposed model that complements consumer consciousness information (i.e., views on fashion) of non-respondents, we need to evaluate whether it is possible to predict it by using the respondents' questionnaire response data and purchase history data of all consumers. In this experiment, all 3,118 respondents are divided into Q test consumers and $3,118 - Q$ learning consumers randomly.

We use both the purchase history data and the questionnaire response data of the $3,118 - Q$ learning consumers for the training data to learn the model, and use only the purchase history data of Q test consumers for the test data to predict the class. Thus, the effectiveness of the proposed model can be verified by checking whether the class of each test consumer is correctly estimated. In this experiment, the number of test consumers Q is set at $Q = 1,000$. Then, the mean absolute error (MAE; denoted by equation (10)) between the correct answer value of the affiliation probability obtained from the questionnaire data for the test consumer $P_{GMM}(z_k|y_q^*)$ and the estimated value of the affiliation probability obtained from the proposed method $P(z_k|u_q)$ is used as the evaluation criterion.

$$\text{MAE} = \frac{\sum_{k=1}^K \sum_{q=1}^Q |P_{GMM}(z_k|y_q^*) - P(z_k|u_q)|}{KQ} \quad (10)$$

The evaluation experiment yields an MAE value of 0.196. Considering the value 0.284 when all $P(z_k|u_q)$, ($k = 1, \dots, K, q = 1, \dots, Q$) are set to 0.2 as the baseline, the proposed method improves about 31% of the MAE value, which demonstrates the effectiveness of our proposed model. The result suggests that clustering with complementing missing data is possible for consumers without questionnaire data using the proposed model.

5. Analysis results

The results of the evaluation experiment in the previous section confirmed that data by non-answering users can be compensated by applying the proposed model. In this section, we show the results of the analysis by applying the proposed model to the entire data provided by the apparel EC site management company A, then we discuss the findings.

5.1 Analysis conditions

The observation period of the purchase history data used for the analysis is from April 1, 2015 to March 31, 2016, which is the same as for the evaluation experiment. The total number of users is $I = 101,501$ (of which 3,118 are respondents), the total number of products is $J = 421,290$, and the total number of purchases is $N = 1,001,901$. The questionnaire survey was conducted between March 17-20, 2016. We used only the 31 questions about “views on fashion” out of the 107 questions in the questionnaire. Furthermore, in the analysis, similar to the evaluation experiment, the number of factors was set at $F = 12$, and the number of latent classes was set at $K = 5$.

5.2 Analysis results and considerations

i) Factor analysis on questionnaire data

First, we obtain the factor loading amount obtained when the factor analysis is performed on the questionnaire data of the respondents. Then, we focused on the survey questions with large absolute values of factor loading amounts given to each factor and interpreted each factor. The results are shown in Table 1.

Table. 1 Features of each factor obtained from the factor loading amount

Factor	Feature of each factor
1	Fashion is a way to express one's personality, express one's own values, and is part of his/her lifestyle.
2	He/she does not mind wearing casual wear at home, takes care of old things, and takes advantage of bargain sales.
3	He/she wants to try new products and products that other people do not yet have ahead of others, and wants to imitate celebrities.
4	Sensitive to trends in fashion, good at adopting trends.
5	Emphasis on country of origin, material, and comfort.
6	He/she feels that quality and function are quite important.
7	He/she has specific favorite brands. Fashion is a change of pace and part of his/her lifestyle.
8	He/she does not worry about wearing clothes that are similar to the people around them. He/she does not follow fashion conventions.
9	Fashion is a part of his/her lifestyle. He/she pays attention to others.
10	He/she wants to attract the attention of the opposite sex, cares about the opinions of sales clerks.
11	He/she is familiar with traditional fashion brands.
12	He/she does not have particular favorite brands, is insensitive to fashion trends and wears old items.

Furthermore, the cumulative contribution ratio of the factor loading amount obtained when applying the factor analysis is presented below in Table 2.

Table. 2 Cumulative contribution ratio obtained from the factor analysis

Factor	1	2	3	4	5	6
Cumulative contribution ratio	0.046	0.082	0.118	0.152	0.182	0.211
Factor	7	8	9	10	11	12
Cumulative contribution ratio	0.239	0.267	0.290	0.312	0.326	0.338

In addition, the scatter diagram of the factor loading amount given for each survey question about factors 1 and 2, which demonstrated the largest contribution rates, is shown in Figure 2 below.

Figure 2 shows that there are many survey questions in the center, and these questions have a weak relevance to factors 1 and 2. Therefore, it cannot be said that all 31 questions can be effectively utilized with these two factors, and thus it is necessary to set more numbers of factors. However, there are multiple points that are further away from the center of Figure 2. These questions have a strong relevance to factors 1 or 2. Therefore, for these factors, the feature of these questions are able to be captured by these factors. Furthermore, for these two factors, the scatter plot of the factor scores given for each respondent is shown in Figure 3 below.

Figure 3 shows that respondents are plotted across a wide range. This shows that there is wide diversity in the opinions of respondents for factor 1 (Fashion is a way to express one's personality, express one's own values, and is part of his/her lifestyle.) and factor 2 (He/she does not mind wearing casual wear at home, takes care of old things, and takes advantage of bargain sales.). However, compared with the right-side, the left-side of Figure 3 has more plots, which means that there are many respondents who do not agree with factor 1. On the other hand, there are many located on the right-side of Figure 3 who agree with factor 1 and have strong interest in fashion.

Figure. 2 Scatter plot of factor load amounts for factors 1 and 2

Figure. 3 Scatter plot of factor scores related to factors 1 and 2

ii) Clustering analysis of respondents by the GMM

Next, we apply the factor scores derived by the factor analysis to the GMM. Based on the probability of each respondent belonging to each class, we calculated the averages of each factor score of respondents belonging to each class. The results are presented in Table 3.

Table. 3 Average factor scores of respondents belonging to each class

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9	Factor10	Factor11	Factor12
z_1	0.043	-0.020	-0.065	0.027	-0.043	0.009	0.001	0.020	-0.068	0.429	-0.065	0.096
z_2	-0.093	0.032	-0.180	-0.278	0.006	-0.002	-0.022	-0.010	0.054	-0.329	-0.122	-0.017
z_3	-0.003	-0.011	0.162	-0.043	0.008	0.002	0.007	-0.026	-0.043	0.028	-0.029	-0.006
z_4	0.009	0.010	0.004	0.005	0.021	-0.001	-0.012	0.014	0.007	-0.013	0.312	-0.002
z_5	0.043	-0.011	0.078	0.288	0.008	-0.008	0.026	0.002	0.050	-0.114	-0.096	-0.071

Table 4 summarizes the features of each class that were observed by the average factor scores of respondents belonging to each class obtained from the GMM.

Table. 4 Characteristics of each class observed by the factor analysis and the parameters of GMM (summary)

z_k	Characteristics of respondents	$P_{GMM}(z_k)$
z_1	Focus on heterosexuals and the opinions of the surrounding people	0.120
z_2	Insensitive to fashion, weak interest in new things	0.673
z_3	Strong interest in what other people do not have	0.026
z_4	Familiar with traditional brands	0.026
z_5	Sensitive to fashion and other trends	0.155

Table 4 shows that as a result of clustering using the questionnaire data provided by the respondents, the characteristics of the consumer consciousness for those in different classes are quite different.

5.2.3 Clustering by the proposed model

We calculated the occurrence probability of each class based on the results of performing clustering on all consumers in the proposed model using the clustering results of the respondents.

Table 5 shows the difference between $P_{GMM}(z_k)$ calculated for only the respondents by GMM and $P(z_k)$ calculated by the proposed model by expanding the target to all consumers. The results show that consumer consciousness trends when clustering is performed using questionnaire data of only the respondents can compensate for the change in consumer consciousness trends when it is expanded to all consumers.

Table. 5 $P(z_k)$ obtained from the proposed model

z_k	z_1	z_2	z_3	z_4	z_5
$P(z_k)$	0.129	0.632	0.038	0.040	0.160

To summarize the above, we applied all data to our proposed method and the results of the clustering are as shown in Table 6.

Table. 6 Characteristics of each class observed by the proposed model and the parameters of the proposed model (summary)

z_k	Characteristics of consumers	$P(z_k)$
z_1	Focus on heterosexuals and the opinions of the surrounding people	0.129
z_2	Insensitive to fashion, weak interest in new things	0.632
z_3	Strong interest in what other people do not have	0.038
z_4	Familiar with traditional brands	0.040
z_5	Sensitive to fashion and other trends	0.160

The occurrence probability of each class and the characteristics of consumers are represented in Table 6 to show the characteristics of those belonging to each class obtained from the proposed model. The top three products presumed to have the highest purchase probability by consumers in each class are shown in Table 7.

In Table 6, we see the characteristics of purchase behaviors for each latent class, and from Table 7 it is possible to interpret the different items estimated to have the highest purchase probability for each class. This result shows that although many popular items appear at the top, different products are ranked higher in each class. For example, class z_3 shows the characteristic of “strong interest in what other people do not have,” and thus we see product categories not seen in other classes such as “watches.” In this way, it appears that the characteristics of each class affects their purchase preferences.

Table. 7 Affiliation probability for each class (top three items)

z_k	Top 1	Top 2	Top 3
z_1	shirt A	tee shirt E	bag B
z_2	shirt F	sweater C	tee shirt E
z_3	chino pants	watch A	tee shirt L
z_4	shirt A	knit A	outerwear B
z_5	shirt A	jacket A	chino pants B

Next, among the top 30 items with the highest occurrence probability for each class, the number of items that were also seen in the top 30 in at least one of the other classes was summarized in Table 8 as shown below.

Table. 8 (In the top 30 items) Number of overlap products with other classes

	z_1	z_2	z_3	z_4	z_5
Number of overlap items	13	12	10	6	16

From this result, we see that consumers belonging to class z_4 (“familiar with traditional brands”) have the least overlap in preferred products with consumers on other classes. In addition, users belonging to class z_3 (“strong interest in what other people do not have”) also have fewer duplicate items. This aligns with our experiential knowledge (i.e., a user who has “low concern for the items held by other people” do not tend to buy popular items, and thus, the number of items overlapping with other classes is small). On the other hand, classes that have the highest number of overlapping items with other classes is class z_5 (“sensitive to fashion and other trends”). From this result, it can be expected that fashion trendy consumers prefer products that tend to be liked by many people.

Table. 9 (In the top 30 items) Average unit price of top preferred items (in Japanese Yen)

	z_1	z_2	z_3	z_4	z_5
Average unit price of items	4,333	3,600	3,766	3,333	3,200

Table 9 shows that there is a difference in the average unit price of the top products preferred by consumers belonging to each class, and we see that class z_1 (“Focus on heterosexuals and the opinions of the surrounding people”) prefers products with higher unit prices than the other classes. In other words, consumers in this class tend to buy expensive products because they care about the opinions of others. In addition, the average unit price of class z_5 (“sensitive to fashion and other trends”) is the lowest. This suggests that consumers in this class prefer to buy inexpensive fashionable clothes.

We conclude that our proposed model makes it possible to reveal the characteristics of the different classes of consumers by clustering by taking into considering both purchase trends and customer consciousness.

6. Discussion

The proposed method can be applied not only to data of this specific apparel EC site, but also to various fields as long as there exists questionnaire data administered only to a subset of consumers and purchase history data for all consumers. Although there are some additional factors to be considered, such as the method to determine the number of factors in the factor analysis, or the method to select questionnaire survey questions, the proposed method can be considered highly versatile.

In addition, the results can be completely different and new insights obtained if we apply the proposed method to the data of the same site by varying questionnaire survey questions or methods to select consumers. In other words, more information can be obtained from this model depending on what the analyst wishes to understand (e.g., by changing the questionnaire survey questions used in the proposed model and applying prior processing steps such as data filtering). Therefore, it is necessary to carefully consider the type of consumer consciousness that should be reflected in the questionnaire survey questions.

Ishigaki et al.(2011)’s model assumes users’ latent classes for users and items’ latent classes for items independently, the relationship between items and users is modeled by combinations of these latent classes. It means there is a constraint that the distribution of items can represent only patterns of items’ latent classes if users’ latent classes are decided. On the other hands, because users and items in the proposed model are linked directly with latent classes, the co-occurrence relationship between users and items can be flexibly expressed without any limitations. From these differences, it is possible to distinguishably use these models according to an actual target problem. However, in Ishigaki et al.(2011)’s model cannot be applied to cases when the questionnaire data are missing for some users. On the other hand, because the proposed model can estimate consciousness of users who did not respond to the questionnaire, we think that the proposed model is significant.

7. Conclusions and future research agenda

In this study, we proposed a new analysis method by analyzing purchase history data and questionnaire data simultaneously. In the proposed method, the purchase trend obtained from purchase history data of all users and consumer consciousness data obtained from a subset of users (i.e., respondents) are used to reveal consumer preference. Specifically, we proposed a model that can supplement missing information by using purchase history data to reflect the consciousness of those that did not answer the questionnaire.

In addition, based on the results of the evaluation experiment, we found that it was possible to complement the missing information from non-respondents to a certain extent by utilizing the proposed model. Thus, the effectiveness of the method was verified. We obtained insights and confirmed the effectiveness of this method by analyzing results obtained from applying the proposed model to actual data. Based on these facts, it became clear that clustering can be performed considering consumer consciousness not only to questionnaire respondents, but to all consumers by applying the proposed method to purchase history data and questionnaire data on a subset of consumers.

In the problem of the proposed method, although the latent classes are assumed between the user and the item and a conditional probability on the user and item sets is defined under a latent class, the occurrence probability of the user with respect to the answering user is calculated by combination of the questionnaire data and the purchase history data. In other words, it is not certain whether the learning procedure is appropriate on the assumption of conditional independence. It means that it cannot be said that the proposed method has rationality from the viewpoint of maximizing likelihood, and improvement of the model considering this point can be said to be a future task.

In addition, with respect to the data actually used this time, we observed that there are no large differences in distributions of users' the demographic attributes, e.g., sex, age, the purchasing time zone and the using device like PC, smartphone etc., between answering and non-answering users. Furthermore, we showed that the customers' awareness of the non-answering users could be accurately acquired for the case that the ratio of the answering and the non-answering users is 2:1. However, it seems to be difficult to estimate the probabilistic relation accurately when there is a large difference in the demographic attributes between answering and non-answering users in practice. And more, if the sample size of the answering user is too small, it is likewise inferred that an accurate estimation is difficult. Since it is necessary to conduct detailed experiments to find out what type of data is difficult to estimate, this should be a future task.

References:

- Abe, M. and Kondo, F. (2005), *Science of marketing - Analysis of POS data* -, Tokyo, Japan.: Asakura Shoten. (in Japanese).
- Data analysis competition (2016), *Joint Association Study Group of Management Science*, <https://jasmac-j.jimdo.com/>
- Dempster, A. P., Laird, N. M., and Rubin, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Society, Series B*, Vol. 39, No. 1, pp. 1-38.
- Goto, M. and Kobayashi, M. (2014) *Introductory pattern recognition and machine learning*, Tokyo, Japan.: Corona Inc. (in Japanese).
- Hirai, Y. (1970) *The first pattern recognition*, Tokyo, Japan.: Morikita Publishing Co., Ltd. (in Japanese).
- Hofmann, T. (1999), "Probabilistic Latent Semantic Analysis," *Proceedings of the Fifteenth conference on Uncertainty in Artificial Intelligence*, pp. 289-296.
- Ishigaki, T. (2010), "Bayesian net user model combining large scale data on purchasing behavior," *The 24th Annual Conference of the Society for Artificial Intelligence* (in Japanese).
- Ishigaki, T., Takenaka T., and Motomura, Y. (2011), "Customer Behavior Prediction by Fusion of Large Scale Data on Daily Behavior," *Journal of Artificial Intelligence Society*, Vol. 26, No. 6, pp 670-681. (in Japanese).
- Ishigaki, T. (2011), "Automatic Extraction Method of Category Bell-dependent Variable Relationships from POS Data with Department Store ID," *Public Interest Corporation Association of Japan Operations Research*, Vol.56, No.2, pp.77 - 83. (in Japanese).
- Nagata, Y. and Munechika, M. (2001) *Multivariate analysis method*, Tokyo, Japan.: Science Co., Ltd. (in Japanese).
- Miyagawa, M. (1987), "EM algorithm and its surroundings," *Applied Statistics*, Vol. 16, No. 1, pp. 1 - 21. (in Japanese).

- Japanese).
- Mizuno, K. (1996) *Multivariate data analysis lecture*, Tokyo, Japan.: Asakura Shoten. (in Japanese).
- Namatame, T. and Suzuki, M. (2013), “Analysis of website browsing behaviors and purchasing behaviors on the EC site,” *The Management Information Society of the General Association*, Tokyo, Japan. (in Japanese).
- Nagata, Y. and Munechika, M. (2001) *Introduction to multivariate analysis method*, Tokyo, Japan.: Science Co., Ltd. (in Japanese).
- Okamoto, T. (1970) *Factor Analysis Method*, Tokyo, Japan.: Union of Japanese Scientists and Engineers. (in Japanese).
- Rasmussen, C. E. (1999), “The Infinite Gaussian Mixture Model,” *NIPS’99 Proceedings of the 12th International Conference on Neural Information Processing Systems*, pp. 554-560.
- Sagawa, M. and Hirooka, Y. (2003) *Marketing / Data Analysis*, Tokyo, Japan.: Asakura Shoten. (in Japanese).
- Shinnou, H. (2007) *Cluster analysis learned by R*, Tokyo, Japan.: Ohmsha/Development Bureau. (in Japanese).
- Shimizu, R., Sakamoto, T., Yamashita, H., and Goto, M. (2017), “Proposal of a Purchase Behavior Analysis Model on EC Site Considering Questionnaire Data,” *Communication of JIMA (Japan Industrial Management Association)*, Vol.27, No.2, pp.70-76. (in Japanese).
- Ueda, S. (2004), “Probabilistic model of multiple topic text - the forefront of text model research -,” *Information Processing Society*, Vol.45, No.3, pp. 282 - 289 (in Japanese).
- Xue, G. R., Wenyan, D., Qiang, Y., Yong, Y. (2008), “Topic-bridged PLSA for cross-domain text classification,” *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. (ACM)*, pp. 627-634.

[DOI : 10.17929/ tq.4.1]

Received: April 6, 2017

Revised: May 18, 2018

Accepted: June 7, 2018