Analytical Model of Customer Purchasing Behavior Considering Event Characteristics on Flower Delivery Business

Aya Kitasato1*, Kenya Nonaka1, Haruka Yamashita2 and Masayuki Goto1

1Waseda University, 3-4-1 Ookubo, Shinjuku-ku, Tokyo, 169-8555, JAPAN
2Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo, 102-8554, JAPAN

ay kitasato@akane.waseda.jp

Abstract:
Currently, a large amount of data accumulated in electronic commerce (EC) sites, and it is possible to customize a marketing strategy for each customer. This study focuses on an analytical model of purchasing data cumulated on a fresh flowers EC site for their marketing analysis. On an EC site providing fresh flower products, the purchase frequency per customer is much lower than that on EC sites selling commodity products or books. Therefore, the accumulated data for each customer on an EC site providing fresh flowers are limited and it is difficult to analyze the individual preferences of each based on their purchasing history. On the other hand, fresh flower products are characterized by their purchase for several events in an individual’s life such as birthday and funeral. Moreover, the selected colors and prices of flowers tend to be different depending on the event. In this research, we focus on the relationship between the attributes of customers and purchasing actions, making it unnecessary to analyze the purchasing behavior of individual customers. We propose the analytical model considering the relationship between the event, the color and price of flower items, and other factors. Because customers’ purchase behaviors are very diverse, the variation in the relationship may be difficult to explain using conventional methods with hard clustering, wherein each datapoint belongs strictly to one class. Therefore, we propose a new latent class model to analyze purchasing behavior data. The proposed model is expected to reveal customers’ purchase behaviors and, thus, help to develop efficient marketing strategies. Moreover, we validate the effectiveness of our proposed method by applying it to the real purchase history data on the EC site.

Keywords
EC Site, Latent Class Model, Purchase History Data, Fresh Flower

1. Introduction
In recent years, the number of consumers who buy various products via the Internet has increased significantly, and a large amount of purchase history data has accumulated on each EC site. Various data analyses have been carried out to propose marketing measures that use these data effectively, and many have been successful in customer analysis and product recommendation (Iwata et. al., 2011; Iwata et. al., 2011). The purchase history data accumulated on the databases are a precious resource for business management and they can be utilized for appropriate purposes. In particular, the data utilization has been highly developed on shopping mall EC sites where the same customers purchase frequently; for example, recommendations are made based on the purchasing tendency and browsing history of each customer. However, various specialty EC sites have characteristics different from those of shopping mall EC sites. Hence, analytics targeting such specialty EC sites should be developed.

In this research, we focus on the case of an EC site providing various flower products online. A target fresh flowers EC site (hereinafter referred to as EC site A) provides direct delivery services for fresh flower merchandise upon receiving orders through the Internet. However, the number of times a customer purchases fresh flowers during the year is extremely small compared to the purchasing frequency of common products such as daily necessities and books. Therefore, on EC site A, the purchase history data for the same customer is limited. In addition, fresh flower products are mostly purchased as gifts on various occasions, which means that the purchased item may depend on the recipient’s personal preference rather than that of the purchaser. This makes it
difficult to analyze the preference of each customer based on their purchasing history on EC site A. Therefore, we focus on the events and attributes of customers. Fresh flower products are purchased in connection with events such as birthdays and Mother’s Day (hereafter, referred to as “events”); therefore, analysis focusing on events is effective. When a customer selects and purchases a product, they reference the category, color, and price of the product. In addition, the purchasing tendency differs depending on customers’ attributes (gender and age).

We, therefore, focus on the item category, color, and price of flowers and the age and gender of customers in each event and construct a statistical model that can analyze customers’ purchasing behavior based on this information. We also consider the diversity of customers to construct an analytical model so that we could introduce a latent class model. The latter assumes discrete hidden variables that cannot be observed directly from the data and analyzes the structure behind the data. It is well-known as a statistical model for analyzing data with multiple characteristics (Magidson and Vermunt, 2004; Ghosh, 2010).

Probabilistic latent semantic analysis (PLSA), proposed by Hofmann (1999), is a typical method of the latent class model. PLSA is a probabilistic model that considers each event as probabilistically arising from a latent class in multiple latent classes. Although PLSA was originally proposed for natural language processing, it can be applied to the purchase history data of customers, i.e., it is possible to express the differences in customer characteristics and purchasing tendencies probabilistically. Shimizu, et al. (2020) applied this model to an analysis of credit card users’ behavior. Said, et al. (2009) studied folksonomy on the Internet. Hyung et. al. (2014) applied this mode to the recommendation of music. In this research, we propose a latent class model that expresses the relationship between event information, customer information, and product information. Moreover, for the parameter estimation, EM algorithm (Hofmann, 1999) is applied. However, when there include the extremely frequent values observed in the data, like stop-words (Li, et al: 2018) in document data, these values are difficult to become characteristics of a latent class. This is because data with these values are too many to be concentrated into a specific latent class and are soft clustered into different latent classes by the characteristics of other variables. Although it is not a significant problem in document analysis, it is desirable to improve this problem on the application of the proposed model to marketing analysis. Therefore, we proposed a simple method to setting the initial value of model parameter for the EM algorithm. By this method, it is possible to acquire a reasonable result of learning for the proposed latent class model. We also present a method for analyzing customer characteristics and purchasing behavior at various events on a fresh flowers EC site. The usefulness of the proposed model is demonstrated by applying the proposed model to actual flower purchase history data on EC site A and clarifying the customer attributes and purchasing tendency for each event. Furthermore, we consider the analysis results and suggest interpretations.

2 Data Description

In the purchase history data handled in this research, there are two types of orders: those from individual customers and corporations. Since this study aims to analyze purchasing behavior based on individual preferences, we exclude purchase history data of corporations and utilize those concerning customers registered as individual members.

Here, we show the basic statistics of the purchase history data on EC site A from August 2018 to September 2019. **Fig. 1** shows the annual purchasing frequency of each customer. **Fig. 2** and **Fig. 3** show the annual sales volume stratified by events and colors of flower items, respectively, from number 1 to number 20 in descending order.

![Fig. 1. Annual purchasing frequency](image-url)
As per Fig. 1, almost no customers purchase flowers multiple times a year. Since most customers purchase them only once a year, it is difficult to estimate each customer’s preferences accurately from their purchase history. Even if these can be estimated, daily promotions based on such detailed information would be meaningless because most customers do not purchase flower products more than once a year. **Fig. 2** shows that Mother’s Day and Birthday account for about 52% of the total annual sales volume. Although this is evident, the percentages of the remaining events vary, and customers have various reasons for purchasing flowers. From these results, it is clear that the accumulated data for each customer on an EC site are limited and customers purchase flowers for various events. Therefore, it is difficult to analyze the preference of each customer, and we should focus on events as mentioned above. **Fig. 3** shows that Mix and Pink account for 70% of the total annual sales volume. Thus, it is evident that most customers do not have a varied preference for the color of flowers.

### 3 Related Work

As mentioned, because of the complex structure of sparse data in a high-dimensional space, simple clustering methods such as K-means, used by MacQueen (1967), or Group average method, used by Larsen and Aone (1999), are insufficient to analyze customers’ preferences. Because the number of items is large, the customer vector is high dimensional and sparse if it is constructed using frequencies of purchased items. For data of an uncertain structure with high dimensionality and sparsity, the latent class model, which is also called the topic model (Magidson and Vermunt, 2004; Ghosh, 2010), is effective.

By assuming the existence of latent variables behind data, each customer’s data are considered to belong to latent classes probabilistically, and the estimated latent class model expresses the relationship between each variable with this model. In marketing analysis, latent class models are sometimes effective because customers’ characteristics are usually not unique and can be classified into certain types. The market consists of several different types of customers and, therefore, the latent class model is a strong tool for analyzing the market. Various studies have been conducted on the latent class model. Mei and ChengXiang (2006) studied document classification related to this model. Iwata (2011) applied this model to the analysis of purchasing history data.

PLSA (Hoffman, 1999) is a method of latent class models that is used to grasp the purchase tendency. The co-occurrence relation of customers and items is expressed with conditional probability distribution in latent classes and Ishigaki (2010) used it to grasp customers’ preferences.

Let the latent classes $Z = \{ z_k : 1 \leq k \leq K \}$ be a set of $N$ items $X = \{ x_n : 1 \leq n \leq N \}$, and a set of $M$ users is denoted by $Y = \{ y_m : 1 \leq m \leq M \}$.

The probabilistic model formula is expressed by Eq. (1).

$$P(x_n, y_m, z_k) = P(x_n | z_k) P(y_m | z_k) P(z_k). \tag{1}$$

A model introducing such latent classes enables modeling that considers latent differences in the purchase situation. Because the latent class $z_k$ cannot be observed as given data, the estimation and maximization algorithm (EM algorithm) was applied by Dempster et. al. (1977) for parameter estimation.
4 Proposal of an Analytical Model

4.1 Concept of Proposal

In this study, we target EC site A, whose business is the delivery of fresh flowers, and our purpose is to analyze the purchase tendency of users on this site. Constructing a model that can analyze customer purchasing behavior is effective. However, as described in section 2, it is evident from focusing on the purchase frequency of each customer on EC site A that most customers only purchase fresh flower products for events such as Mother’s Day and Birthdays, making purchases through the site only once or twice a year. Therefore, the purchase history data for each customer are very limited, and it is relatively difficult to analyze the purchase behavior that captures the preferences of individual customers compared to other EC sites that handle daily necessities and books. In this study, we focused on the relationship between the attributes of customers. In addition, since it is naturally assumed that each customer takes a purchasing action based on the characteristics of fresh flowers (use, color, etc.), we study, we focused on the relationship of these pieces of information. We assumed a latent relationship between multiple variables.

In the proposed latent class model, the co-occurrence of a customer’s gender, age, use of fresh flowers (event information), color, unit price, and item category is assumed to be conditional independent of one another on a given latent class.

4.2 Formulation of the Proposed Model

The co-occurrence relation of age and gender as characteristics of customers and item category, color, and unit price as characteristics of flowers is expressed. Thus, it is possible to determine the tendency to purchase flowers, such as the type of event, customer, and fresh flowers, using our model. We denote a finite number of events $U = \{u_i; 1 \leq i \leq I\}$, a set of gender $E = \{e_j; 1 \leq j \leq J\}$, a set of age $A = \{a_l; 1 \leq l \leq L\}$, a set of colors $O = \{o_c; 1 \leq c \leq C\}$, a set of categories $G = \{g_m; 1 \leq m \leq M\}$, $t$ as a unit price of flower since the unit price assumes a continuous value, and a set of $Z = \{z_k; 1 \leq k \leq K\}$. At this time, the proposed probabilistic model formula with the latent class $z_k$ that is not observed is expressed by Eq. (2).

$$P(u_i, e_j, a_l, o_c, g_m, t, z_k) = P(u_i | z_k)P(e_j | z_k)P(a_l | z_k)P(o_c | z_k)P(g_m | z_k)P(t | z_k)P(z_k). \tag{2}$$

Fig. 4. Graphical model of proposed model

We assumed a multinomial distribution for the probability distribution of event, age, gender, color, and category $P(u_i | z_k), P(e_j | z_k), P(a_l | z_k), P(o_c | z_k)$, and $P(g_m | z_k)$; a multinomial distribution for $z_k$; and a normal distribution expressed by Eq. (3) for the unit price of each flower.

$$P(t | z_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{(t - \mu_k)^2}{2\sigma_k^2}\right\}, \tag{3}$$

where $\mu_k$ and $\sigma_k$ represent the mean and variance of the $k$-th normal distribution, respectively.

4.3 Parameter Estimation By EM-algorithm

In all $N$ purchase history data of EC site A, the $n$-th event is denoted by $u_n \in U$, the gender of a user is $s_n \in E$, the age of a user is $h_n \in A$, the color of the flower is, $w_n \in O$ and the category of flower is $r_n \in G$ and the unit price is $p_n$, then the $n$th purchase data can be represented by these co-occurrences...
\( (b_n, s_n, h_n, w_n, r_n, p_n) \). Then, the log-likelihood \( LL \) for all \( N \) purchase history data is expressed by Eq. (4) below. In the model formulations such as Eq. (2), the model is described using a continuous variable \( t \), whereas the training data is written as \( p_n \). The notation of price in the training data is changed corresponding with that for other categorical variables. Other features in Eq. (3) are categorical variables, for example, events are represented as \( u_i \) for the type of event, and gender as \( e_j \) for the type of gender. When calculating the occurrence probability for each discrete value, the argument \( i \) of \( u_i \) means the number of events whereas the argument \( n \) of \( b_n \) means the number of samples. If we use the same notation with different arguments, \( u_i \) and \( u_n \), these will lead to confusion. Therefore, all the notations including the other features are changed as the values at the \( n \)-th data, such as \( b_n \) for events and \( s_n \) for gender. The unit price is also changed to \( p_n \) for the training data although the value is uniquely determined even if it is set as the value of \( n \)-th data like \( r_n \). Therefore, when the calculation of the likelihood in equation (4), the notation is changed to \( p_n \); however, the normal distribution of unit price is assumed that

\[
P(p_n | z_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(p_n - \mu_k)^2}{2\sigma_k^2}\right).
\]

The parameters of this model are estimated using the EM algorithm \([14]\) based on the following updating formula to locally maximize the log-likelihood function \( LL \).

\[ LL = \sum_{n=1}^{N} \sum_{k=1}^{K} \log P(b_n, s_n, h_n, w_n, r_n, p_n, z_k). \tag{4} \]

**E-step)**

\[
\gamma_{nk} = P(b_n | z_k)P(u_n | z_k)P(h_n | z_k)P(w_n | z_k)P(r_n | z_k)P(p_n | z_k)P(z_k), \tag{5}
\]

\[
P(z_k | b_n, s_n, h_n, w_n, r_n, p_n) = \frac{\gamma_{nk}}{\sum_{k=1}^{K} \gamma_{nk}}. \tag{6}
\]

**M-step)**

\[
P(z_k) = \frac{1}{N} \sum_{n=1}^{N} P(z_k | b_n, s_n, h_n, w_n, r_n, p_n), \tag{7}
\]

\[
P(u_i | z_k) = \frac{1}{NP(z_k)} \sum_{n=1}^{N} \delta(b_n, u_i)P(z_k | b_n, s_n, h_n, w_n, r_n, p_n), \tag{8}
\]

\[
P(e_j | z_k) = \frac{1}{NP(z_k)} \sum_{n=1}^{N} \delta(s_n, e_j)P(z_k | b_n, s_n, h_n, w_n, r_n, p_n), \tag{9}
\]

\[
P(a_i | z_k) = \frac{1}{NP(z_k)} \sum_{n=1}^{N} \delta(h_n, a_i)P(z_k | b_n, s_n, h_n, w_n, r_n, p_n), \tag{10}
\]

\[
P(o_j | z_k) = \frac{1}{NP(z_k)} \sum_{n=1}^{N} \delta(w_n, o_j)P(z_k | b_n, s_n, h_n, w_n, r_n, p_n), \tag{11}
\]

\[
P(g_m | z_k) = \frac{1}{NP(z_k)} \sum_{n=1}^{N} \delta(r_n, g_m)P(z_k | b_n, s_n, h_n, w_n, r_n, p_n), \tag{12}
\]

\[
\mu_k = \frac{\sum_{n=1}^{N} P(z_k | b_n, s_n, h_n, w_n, r_n, p_n)p_n}{\sum_{n=1}^{N} P(z_k | b_n, s_n, h_n, w_n, r_n, p_n)}. \tag{13}
\]
where \( \delta(\alpha, \beta) \) is an indicator function that takes 1 when \( \alpha = \beta \) and in other cases \( \delta(\alpha, \beta) = 0 \).

In this study, we deal with the large sized data including several categorical variables taking discrete values. For example, the variable “Event (\( u_i \))” takes a value in a discrete set with 40 elements meaning events (\( i = 1, \ldots, 40 \)) such as “birthday” or “Mother’s Day”. If we summarize the categorical data using a contingency table (Ku and Kullback, 1974) by ignoring the continuous variable “Price of the flower (\( t \))”, the number of cells in the table becomes \( 40 \times 2 \times 12 \times 45 \times 148 \), and the number is so larger than sample size; thus, the data is so sparse.

The statistical model in this study is different from the usual categorical data analysis based on a contingency table proposed by Ku and Kullback (1974). The model assumes the conditional independence between observed variables under the latent classes; therefore, probabilities of each pattern of all combinations of possible discrete variables’ values can be calculated by \( \sum_k \, P(z_k | u_i) \, P(e_j | z_k) \, P(\alpha_l | z_k) \, P(g_m | z_k) \, P(t | z_k) \) for all \( u_i, e_j, \alpha_l, g_m, t \). We can represent any of all probabilities on the data space by using the parameters, \( P(u_i | z_k), P(e_j | z_k), P(\alpha_l | z_k), P(g_m | z_k), P(t | z_k) \), and the number of the parameters is \( (k-1) + k \times ((40-1) + (2-1) + (12-1) + (45-1) + (148-1)) \) \( + k \times 2 \). This model enables to reduce the number of parameters greatly compared with a model assuming a contingency table. It is, therefore, possible to estimate the probabilistic structure with high accuracy by observed sample data. In addition, the sample space considered by the proposed model is rather less sparse than the traditional topic models such as PLSA or LDA. This is because PLSA and LDA assume probability events for all individual users while the proposed model represents it by user attributes and events.

Moreover, it is desirable to consider that “birthdays” and “mother's days” in the event variable account for 80-90% of the total purchasing records in this study and that there is a large bias in the training data set. This problem happens in the various data analysis such as the text data analysis which is called stop-words problem (Li, et al.: 2018). Usually, stop words are eliminated from the data in the data cleaning phase for text data analysis, and this approach decreases the number of the data and leads to the estimation problem. However, the events of “birthdays” and “mother's days” are important values in this analysis so that they cannot be eliminated. For the model, we propose a simple method, that is, devising the settings of the initial value on the EM algorithm for a latent class model. The proposed method is to set the initial values of the conditional probability under the latent class as 1.0 for the frequent patterns of one of the variables. Here is one of the impacts of the proposed method.

For this study, “birthdays” and “mother's days” of \( P(u_i | z_k) \) are majority in the whole data. Therefore, we set the conditional probability of the two realizations to be 1. This enables to extract the latent class that are easy to give the interpretations. Note that, the devise can be applied to many data analysis including the problem of stop-words.

5. Actual Data Analysis

5.1 Experimental Conditions

We apply the proposed model to actual purchase history data of fresh flower products on EC site A, which are provided by EC site A through the database system. Since the raw data are provided only for the joint research purposes and are not available to be opened, only the analysis results are described without showing the raw data, same as the many conventional studies on business data analysis (e.g. Xu et al., 2013; and Saker, 2019). The data size used in this analysis was about 400 thousand samples. The exact number cannot be opened in this paper because of the request by the company providing the data for the purpose of academic works. 

The data were accumulated from August 2018 to September 2019. The number of event types was \( I = 40 \), the number of gender types of customers was \( J = 2 \), the number of age types of customers was \( L = 12 \), the number of color types of the flower was \( C = 45 \), and the number of category types was \( M = 148 \). Furthermore, the number of latent classes in the latent class model was set to \( K = 12 \). This was chosen from the viewpoints of both the AIC value and the least overlap of class features in the interpretation as follows:

First, we varied the number of classes up to \( K = 14 \), changed the initial values 20 times in all \( K \), and obtained the average of the 20 patterns of AICs for each class. The results are shown in Table 1. This table shows that the average value of AIC is preferable when the \( K = 5 \), 8, 12. Note that AIC is an information criterion derived for model classes whose maximum likelihood estimator follows asymptotically normal distribution, a latent class model usually does not satisfy the assumption of asymptotic normality. In addition, there is no
guarantee for the EM algorithm to search the strictly maximum likelihood estimator for the latent class model. Therefore, the value of AIC cannot be an absolute criterion for deciding the number of latent classes but can be used as a reference value. In this study, we compared the interpretability of estimated latent classes for each number of \( K \) in addition to check the AIC values. When \( K = 12 \), the differences of characteristics between latent classes were obvious and the average value of AIC is enough low, so we decided the number of latent classes as \( K = 12 \).

In this case, the classes with \( P(u_i|z_{12}) \) initially set to 1.0 for Birthday and Mother’s Day are \( z_1 \) and \( z_2 \), respectively.

5.2 Experimental Result

The analysis results obtained by employing the proposed model are shown in Tables 2 to 5. Table 2 shows the estimated probability of each latent class \( P(z_k) \). The latent class \( z_1 \) had the highest probability, \( P(z_k) \) and from \( z_1 \) to \( z_{12} \), the latent classes were sorted in descending order. The results that estimated conditional probabilities of characteristics of users (except gender) and flower items under \( z_k \) are more than 0.05, and the top three are shown in Table 3 and Table 4. In these tables, we describe the interpretation of the latent classes based on the estimated conditional probabilities. In the results of gender, we show both of gender like “Male and Female” when occurrence of men under \( z_k \) as \( P(a_1|z_k) \) is 0.45 \( \leq \) \( P(a_1|z_k) \leq 0.55 \).

### Table 1. The mean and SD of AIC for each class

| Class | The mean of AIC | The SD of AIC |
|-------|----------------|--------------|
| 3     | 17358662.6     | 232979.8     |
| 4     | 17341490.7     | 210714.7     |
| 5     | 17321788.5     | 209253.9     |
| 6     | 17305231.5     | 199332.8     |
| 7     | 17449103.5     | 173139.4     |
| 8     | 17321618.6     | 197272.8     |
| 9     | 17360547.7     | 214794.3     |
| 10    | 17350053.7     | 255676.0     |
| 11    | 17418627.3     | 185301.1     |
| 12    | 17318456.5     | 213107.4     |
| 13    | 17332361.3     | 199193.1     |
| 14    | 17340273.1     | 144053.8     |

### Table 2. \( P(z_k) \) of each class

| Latent class | \( P(z_k) \) | Latent class | \( P(z_k) \) |
|--------------|-------------|--------------|-------------|
| \( z_1 \)    | 0.193       | \( z_7 \)    | 0.0617      |
| \( z_2 \)    | 0.143       | \( z_8 \)    | 0.0613      |
| \( z_3 \)    | 0.142       | \( z_9 \)    | 0.0408      |
| \( z_4 \)    | 0.131       | \( z_{10} \) | 0.0355      |
| \( z_5 \)    | 0.0773      | \( z_{11} \) | 0.0319      |
| \( z_6 \)    | 0.0622      | \( z_{12} \) | 0.0205      |

The estimated probabilities \( P(z_k) \) in Table 1 indicate the occurrence possibilities of purchases belonging to that class. Because flower items sell well on specific events, such as Mother’ Day, the probabilities of latent classes tend to be biased on a few latent classes.
the relationship between these features related to flower products. This result shows that the proposed model can represent the relationship between events, product attributes (color, category, and average price), and customer attributes (age and gender) from the viewpoint of the impact of marketing measures. That is, we may consider the possibility of appropriate marketing measures for each latent class in descending order by considering its characteristics and the relationship between these features related to flower products.

Table 5 shows the summary of Table 3 and Table 4 and the average unit price of each class. Through Table 5, it is possible to grasp the characteristics of each latent class and the relationship between events, product attributes (color, category, and average price), and customer attributes (age and gender). Since the latent classes are listed in descending order of the estimated probability, latent classes with small values are relatively important from the viewpoint of the impact of marketing measures. That is, we may consider the possibility of appropriate marketing measures for each latent class in descending order by considering its characteristics and the relationship between events, product attributes (color, category, and average price), and customer attributes (age and gender).
between events, product attributes, and customer attributes. However, even if the estimated probability is relatively small, the latent class is important for marketing activity only when the average price is high or the event is important for the site.

### Table 5. The feature of each class

| Latent class | The feature of each class |
|--------------|--------------------------|
| $z_1$ | Female customer, who is in her 40s or 30s, likes the flower which is 6833 yen and pink or white for Offering Flowers and Celebration for opening. |
| $z_2$ | Male customer, who is in his 50s or 40s, likes the flower which is 3929 yen and pink or red and orange for Birthday or Father’s Day. |
| $z_3$ | Female customer, who is in her 30s or 40s, likes the flower which is 3910 yen and pink or red for Mother’s Day. |
| $z_4$ | Male customer, who is in his 50s or 40s, likes the flower which is 4493 yen and pink or red for Mother’s Day or Birthday or Offering Flower. |
| $z_5$ | Male customer, who is in his 40s or 50s, likes the flower which is 4793 yen and yellow and orange or pink for Wedding Anniversary or Birthday or Father’s Day. |
| $z_6$ | Female customer, who is in her 40s or 30s, likes the flower which is 3835 yen and yellow and orange or pink for Birthday. |
| $z_7$ | Female customer, who is in her 40s or 30s, likes the flower which is 4149 yen and mix or pink for Mother’s Day, Celebration for Retirement, Birthday. |
| $z_8$ | Female customer, who is in her 50s or 30s, likes the flower which is 4295 yen and mix or pink for Offering Flowers or Grandparents’ Day. |
| $z_9$ | Male customer, who is in his 50-30s, likes the flower which is 6429 yen and yellow and orange or white for Birthday or Mother’s Day. |
| $z_{10}$ | Male customer, who is in his 40s or 50s, likes the flower which is 4162 yen and yellow and orange or red for Mother’s Day or Grandparents’ Day. |
| $z_{11}$ | Female customer, who is in her 50-30s, likes the flower which is 3719 yen and mix or yellow and orange for Celebration for Birthday or Father’s Day. |
| $z_{12}$ | Female customer, who is in her late 50-30s, likes the flower which is 4298 yen and mix or pink for Offering Flowers and Mother’s Day. |

### 6 Discussions

#### 6.1 Effectiveness of the Proposed Model

To suggest the effectiveness of the proposed model, we interpreted the results from the following two viewpoints and compared them with our experimental knowledge. The latent class $z_1$ has a high probability of occurrence for ‘Celebration for opening of the store’ and the average unit price was 6,833 yen and it is more expensive than other classes. Therefore, it was inferred that the result of $z_1$ matches our experimental knowledge that when the celebration for someone, sending expensive Phalaenopsis and flowers on stand are general in Japan, and it can be thought that our model is reliable. Moreover, it can be suggested that this model could learn other event except Mother’s Day and Birthday because it could be seen that other events occurred in some classes from $z_1$, $z_5$, $z_8$, $z_{12}$ in Table 3. For the above reason, the effectiveness of this model to analyze purchase behavior of customers should be suggested.

#### 6.2 Consideration in Terms of Age of the Consumer

Among the events that occurred this time, it is found from Table 3 that there was not much difference in product preference by age. In the case of $z_9$, $z_{11}$, and $z_{12}$, from Table 3 and 4, a wide range of generations such as 50-30s were involved in the event. It was found that these customers preferred neutral colors such as yellow and orange, which are not the colors that are generally difficult to give depending on gender, such as blue and pink. However, from these results, it is not necessary to change the recommendation of products depending on gender for the events that occurred in this study, if we target a wide range of generations, it can be considered to recommend flowers in
6.3 Consideration in Terms of Other Features of the Data

From Table 3, 4 and 5, it is thought that Mother’s Day and birthdays occur at a high rate in \( z_2, z_3, z_4, z_5, z_6, z_7, z_9, z_{10}, z_{11} \) and \( z_{12} \). Looking at gender and price range in this class, females often purchase items in the 3,000 yen range, while males purchase items in the 4,000 yen range or higher, indicating that males are likely to purchase more expensive items in these events. Therefore, this fact suggests that males can be recommended higher priced flower arrangements at these events.

In addition, for the customer attributes that purchased at birthdays and Mother’s Day, which are occurring in these classes, we can recommend products similar to the fresh flowers purchased at these events for other events such as Father’s Day and wedding anniversaries. These interpretations are difficult to be obtained through the simple statistical approaches, therefore, the results show the adequacy of this research. Furthermore, it can be seen from Table 1 that \( z_{12} \) has the lowest \( P(z_k) \) which is 0.0205; however, considering the purchase history, it is considered that appealing to customers belonging to this class is meaningful enough because there are 8,000 purchases in purchase history data.

7. Conclusion and Future Works

In this research, we formulated a new latent class model to analyze purchase history data, which had a complex structure. We applied the proposed model to the real purchase history data provided by EC site A and analyzed the estimated parameters from the marketing viewpoint. The proposed latent class model represents the co-occurrence relationship between the event, category, color, and unit price of flowers, and the gender and age of customers.

This study has three novelties. First novelty is to formulate the structure of model appropriate for the target problem and to derive the estimation algorithm based on EM algorithm. The model structure is different from the conventional PLSA which only deals with the relationship between items and customers (Hoffman, 1999). That is, the structure of the proposed model is that each purchase information related with the events, the item characteristics, and customer attributes are conditionally independent under the latent classes. Assuming the appropriate distributions for each variable, then we formulated the estimation algorithm for the constructed probabilistic model. The second novelty is to clarify the availability and effectiveness of the proposed model in terms of the interpretability for actual data accumulated in a company by setting the appropriate parameters or distribution of the parameters. Even for similar events, we were able to understand differences in preferences or purchasing tendencies between gender and age. By utilizing these results, it is expected that different products can be recommended for each segment given by gender and age for each event, which may lead to improved customer satisfaction and increased sales volume. The last one of the novelties is the initial setting of the EM algorithm. This enables the improvement of data analysis including the realized values with too many observations. This is the significant improvement for the application to marketing analysis because the characteristics of latent classes become easy to be interpreted.

However, as per the study of Scammon and Shaw (1982), in this EC site, there are also many customers who do not want to take up their time to select a flower item and make a purchase quickly. Therefore, it is more efficient to focus on customers who take time to buy flowers at the EC site to know the purchasing behavior that captures customers’ preferences. Future studies can incorporate the analysis of customer characteristics in detail by utilizing the site browsing history data until each customer makes a purchase to know their motivation. In addition, adding the site browsing history data to the current model will help understand the customers’ preferences in more detail, as in the personal recommendation model proposed by Wu and Ren (2016). Although the effectiveness of the proposed model was clarified by using an actual data provided by a flower arrangement e-commerce site as a target case study, it is necessary to confirm the generality of the proposed method and apply to other cases as a future study.

References

Dempster, A. P., Laird, N. M., and Bubin, D. B. (1977), “Maximum likelihood from incomplete data via the EM algorithm”, Journal of the Royal Statistical Society: Series B (Methodological), Vol.39, No.1, pp.1-22.

Ghosh, J. K. (2010), “Latent class and latent transition analysis: with applications in the social, behavioral, and economic sciences”, Springer, New York.

[DOI : 10.17929/ tqs.7.125] Copyright © 2021 Journal of the Japanese Society for Quality Control. All rights reserved.
health sciences by Linda M. Collins, Stephanie T. Lanza”, *International Statistical Review*, Vol.78, No.3, pp.449-450.

Hofmann, T. (1999), “Probabilistic latent semantic indexing”, *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.50-57.

Hyung, Z., Lee, K., and Lee, K. (2014), “Music recommendation using text analysis on song requests to radio stations” *Expert Systems with Applications*, Vol.41, No.4, pp.2608-2618.

Ishigaki, T. (2010), “Bayesian net user model combining large scale data on purchasing behavior”, *24th Annual Conference of the Society for Artificial Intelligence* (in Japanese), pp.55-57.

Ishigaki, T., Takenaka, T., and Motomura, Y. (2011), “Customer behavior prediction system by large scale data fusion in a retail service”, *Transactions of the Japanese Society for Artificial Intelligence*, Vol.26, No.6, pp.670-681.

Iwata, T., Watanabe, S., and Sawada, H. (2011), “Fashion coordinates recommender system using photographs from fashion magazines”, *Twenty-Second International Joint Conference on Artificial Intelligence*.

Ku, H. H., and Kullback, S. (1974). “Loglinear models in contingency table analysis.”, *The American Statistician*, Vol.28, No.4, pp.115-122.

Larsen, B. and Aone, C. (1999), “Fast and effective text mining using linear-time document clustering”, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.16-22.

Li, C., Lu, Y., Wu, J., Zhang, Y., Xia, Z., Wang, T., and Guo, J. (2018). LDA meets Word2Vec: a novel model for academic abstract clustering, *In Companion proceedings of the web conference 2018*, pp. 1699-1706.

MacQueen, J. (1967), “Some methods for classification and analysis of multivariate observations”, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol.1, No. 14, pp.281-297.

Magidson, J., and Vermunt, J. K. (2004), “Latent class models”, *The Sage handbook of quantitative methodology for the social sciences*, pp.175-198.

Mei, Q. and ChengXiang, Z. (2006), “A mixture model for contextual text mining.”, *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Sakar, C. O., Polat, S. O., Katicircioglu, M., and Kastro, Y. (2019), “Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks”, *Neural Computing and Applications*, Vol.31, No.10, pp.6893-6908.

Scammon, D. E. and Shaw, R. T. (1982), “Is a gift always a gift? An investigation of flower purchasing behavior across situations”, *ACR North American Advances*.

Shimizu, R. and Yamashita, H., Ueda, M., Tanaka, R., Tachibana, T., and Goto, M. (2020), “Latent variable models for integrated analysis of credit and point usage history data on rewards credit card system”, *International Business Research*, Vol.13, No.3, pp.106-106.

Takacs, A., Watzeck, R., Umbrath, W., and Hennig, L. (2009), “A hybrid PLSA approach for warmer cold start in folksonomy recommendation” *Recommender Systems & the Social Web*, pp.10-13.

Wu, S. and Ren, W. (2016), “Personal recommendation using deep recurrent neural networks in NetEase”, *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pp.1218-1229.

Xu, C., Du, C., and Yu, S. (2013), “A novel model for user clicks identification based on hidden semi-Markov”, *Journal of Network and Computer Applications*, Vol.36, No.2, pp.791-798.

Acknowledgement

We extend my sincere gratitude to Hana-cupid Co., Ltd. for their cooperation from the initial stages of this work. This work was supported by JSPS KAKENHI Grant Number 21H04600.

Authors’ biographical notes

**Aya Kitasato** received her B.S. degree in Chemistry from Keio University, Tokyo, Japan in 2019. She is currently pursuing her master’s degree in Industrial and Management System Engineering from Waseda University, Tokyo, Japan. Her research interests include machine learning theory.

**Kenya Nonaka** received his B.L. degree in faculty of law from Hitotsubashi University, Tokyo, Japan, in 2018. He is currently pursuing his master’s degree in Industrial and Management System Engineering from Waseda
Haruka Yamashita is an assistant professor at the School of Science and Technology, Sophia University, Japan. She was born in Tokyo, Japan, on September 30, 1987. She received her bachelor’s degree from Tokyo University of Science and Technology in 2010. Moreover, she received her master’s degree in 2012 and PhD. degree from Keio University in 2015. She was a research associate at Waseda University from 2015 to 2017. Her research interests include applied statistics, multivariate analysis, and machine learning. She is a member of the Japanese Society of Applied Statistics, Japanese Industrial Management Association, and Japanese Society of Quality Control.

Masayuki Goto received his B. E. and M.E. degrees from the Musashi Institute of Technology in 1992 and 1994, respectively. He received Dr. E degree in Industrial and Management System Engineering from Waseda University in 2000. Since 2011, he has been a professor at Waseda University. He is researching in the fields of applied information mathematics, business analytics, data science, and machine learning and its applications. He is a member of IEEE, INFORMS, Information Processing Society of Japan, The Institute of Electronics, Information and Communication Engineers, Japanese Society for Artificial Intelligence, Japanese Industrial Management Association, and Japanese Society of Quality Control.