Modeling customer purchase behavior based on page transitions by latent class model for customer segmentation

Yuki Matsuzaki†
Graduate School of Creative Science and Engineering
Waseda University, Tokyo, Japan.
Tel: (+81) 3-5286-3290, Email: y_matsuzaki@toki.waseda.jp

Kenta Mikawa
Department of Information Science,
Shonan Institute of Technology, Kanagawa, Japan
Tel: (+81) 4-6630-0212, Email: mikawa@info.shonan-it.ac.jp

Masayuki Goto
School of Creative Science and Engineering
Waseda University, Tokyo, Japan.
Tel: (+81) 3-5286-3290, Email: masagoto@waseda.jp

Abstract. Recently, it has become popular to purchase product items through E-commerce sites (EC sites), and the internet market scale has been expanding. Under this situation, many EC sites conduct various kinds of sales promotion activities by analyzing huge amount of customers' purchase histories, and customer segmentation is one of the most important tools in marketing. Particularly, modeling of customers purchase behavior based on probabilistic models such as the Aspect Model (AM) is an attractive way for customer segmentation. The AM focuses on pairs of a customer and an item and it assumes unobserved features such as customers' heterogeneity and items' similarity as latent classes. Although the original AM focuses on pairs of a customer and an item mainly, the data about customers' browsing histories are also available on EC sites. If the model can take in the information of page transitions, it becomes possible to model customers' purchase behavior in detail and make better customer segments. In this paper, we propose a new latent class model that integrates browsing histories in addition to purchase histories by assuming that customers' page transitions can be described by a Markov process. An analysis of actual EC site data is demonstrated to clarify the effectiveness of our proposed model.

Keywords: customer segmentation, marketing, latent class model, Markov model

1. INTRODUCTION

It has become popular to purchase product items through E-commerce sites (EC sites), and the market scale of that has been expanding (Burt and Sparks, 2003). Under this situation, many EC sites conduct various kinds of sales promotion activities by analyzing huge amount of customers' purchase histories. When conducting some promotion activities, it is necessary to consider that what kind of activities should be conducted for what kind of groups of customers. That is, in order to take an appropriate measure to appropriate groups of customers, it needs to be narrowed down a target group. This is because customer segmentation is recognized as one of the most important tools in marketing (Wedel and Kamakura, 2012; Plant, 2000). Concerning customer segmentation, the effectiveness of probabilistic models such as the Aspect Model (AM) (Thorsten and Francine, 2002; Bhatnagar and Ghose, 2004; Bassi, 2007) is widely known. This is because the latent class outputs a descriptive interpretation, reducing the relationships between the observed variables into a relatively small number of latent classes. The AM mainly focuses on relationships between a customer and an item and it assumes unobserved features such as customers' heterogeneity and items' similarity as latent classes (Hofmann, 2004; Yamagami et al. 2014; Matsuzaki et al., 2015).

Although the original AM focuses on pairs of a customer and an item mainly, various kinds of the histories data about customers' browsing activities are also available on EC sites with the recent growth of information technology. These customers' browsing histories strongly reflect the features of customers' purchasing behaviors because this information indicate the processes for purchase or not. Therefore, if the model can take in the information of page transitions, it becomes possible to model customers' purchase behavior in detail and make better customer segments which are useful for finding important targets.

In this paper, we propose a new latent class model that integrates browsing records and purchase records by assuming that customers' page transitions can be described by a Markov process. For constructing the proposed model, we extend the latent segment Markov chain model (LSMC) (Dias and Vurmunt, 2007) that assumes latent classes for page transitions on a web news site. In addition, in order to demonstrate the usefulness of the proposed model, we conduct the experiment using data of a major EC site by using proposed model and analyze the estimated parameters in detail.

2. RELATED MODELS

2.1. The Aspect Model

The Aspect Model (AM) (Hofmann, 1999) is widely known as one of latent class models, and can be applied to many fields, for example, text mining (Mei and Zhai, 2006), user clustering (Ishigaki et al., 2011) and recommender system (Hofmann, 2003). In the AM, it is possible to take account of customers' preferences and items' similarities by introducing an unobserved class variable $z \in \mathcal{Z} = \{z_1, z_2, \cdots, z_Z\}$. Let a set of customer u and a purchased item b be$u \in \mathcal{U} = \{u_1, u_2, \cdots, u_U\}$ and $b \in \mathcal{B} = \{b_1, b_2, \cdots, b_B\}$ respectively. The probabilistic model of AM is the formulated by

$$P(z) = \sum_{z \in Z} P(z)P(u|z)P(b|z) \qquad (1)$$

Although the AM was proposed as a document model, it has been applied to various problems. Latent class models have been especially developed in the research field of marketing such as recommender systems and customer segmentation and their effectiveness has been shown in many studies (Jin et al., 2003; Jin et al., 2006). In these latent class models, the probabilities of a user and an item are conditionally independent each other if a latent class is given. In addition, each user and item are permitted to belong to different latent classes at each event. The graphical model of the AM is described in the Figure 1.

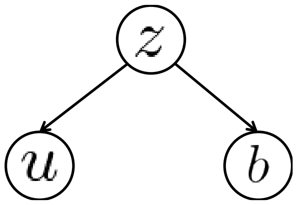

Figure 1. The graphical model of the AM

2.2. The Latent Segment Markov Chain

For the data of page transitions, Dias and Vurmunt proposed a new latent class model based on the AM called the LSMC. The LSMC is based on the assumption of a first order Markov chain for a sequence of page transitions which is called "a session" in order to consider the length difference of each browsing record. Here, let a set of session and a set of page type be defined as $\mathcal{S} = \{s_i : 1 \le i \le I\}$ and $\mathcal{K} = \{k_j : 1 \le j \le J\}$, respectively. In addition, $x_t^i \in \mathcal{K}$ indicates the t th browsed page type of s_i. A session s_i corresponds to browsing pages, $\boldsymbol{x}_i = (x_0^i, x_1^i, \cdots, x_{T_i}^i)$ which consists of T_i length series of page types. Here, assuming the first order Markov chain, *i.e.* the t th browsed page type only depends on the $t-1$ th browsed page type, the probability of s_i's browsing records, x_i is defined as follows:

$$P(\boldsymbol{x}) = P(x_0^i) \prod_{t=1}^{T_i} P(x_t^i | x_{t-1}^i) \qquad (2)$$

For this popular first order Markov property, Dias and Vurmunt introduced an unobserved latent variable $v_i \in \mathcal{Z} = \{z_1 : 1 \le i \le L\}$ which corresponds to a session s_i. In the LSMC, the Equation (2) is extended as the Equation (3) by assuming that each session is clustered into L latent classes, and the Figure 2 shows the graphical model of the LSMC.

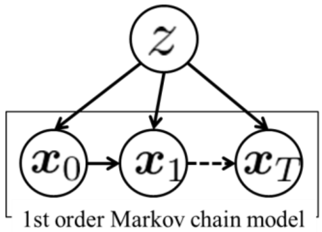

Figure 2. The graphical model of the LSMC

$$P(\boldsymbol{x}|v_i) = P\left(x_0^i|v_i\right)\prod_{t=1}^{T_i} P(x_t^i|x_{t-1}^i, v_i) \qquad (3)$$

In the Equation (3), $P\left(x_0^i|v_i\right)$ indicates an initial distribution of a session s_i which belongs to latent class v_i and $P\left(x_t^i|x_{t-1}^i, v_i\right)$ means the probability that page type x_t^i is browsed at the t th event in a session s_i on condition that x_{t-1}^i is browsed at the $t-1$ th event.

3. PRELIMINARY DATA ANALYSIS

3.1. Data Information

In this paper, we analyze the access log data of a Japanese major EC site as a case study of the application of our proposed model. The session IDs are allocated for each browsing behavior which is started from visiting the EC site and ended by leaving from the site, and each session ID corresponds to series of browsed page type(s), the number of browsing page types, browsing time per page, and a purchase record. From this data, we especially focus on the data of time series of browsed page type(s) and purchase record of sessions. Besides, the demographic data such as sex, age, and occupation are not used for analysis because of the privacy protection policy. In addition, although the data contains the way to access by three devices which are a personal computer (PC), a smart phone, and a tablet, it is better to focus on one device because the site structures among devices are different. In this study, we target the access from PC because of the data size of the access log.

We especially focus on browsing records which consist of time series of browsed page types such as item page and category page. The page types consist of 12 types which are defined by the EC site as mentioned in the Table 2 and 3. We assume that browsing records are characterized by time series of browsed page type(s) and investigate them. The period covered by the dataset is from April 1, 2016 to April 7, 2016, and the data consists of 386,671 sessions and 6,031,916 browsed pages.

3.2. Analysis by the LSMC

Using browsing records of each session, we conduct the preliminary analysis for those data by the LSMC. The results are shown in the Table 1, 2 and 3, and the Table 1 shows features of two latent classes that the one has the highest purchase rate and the other has the lowest. In the Table 2 and 3 which are called Markov maps, each map shows the transition probability from the page type k_j to the page type k_m such as $P\left(x_t = k_j|x_{t-1} = k_m, v\right)$. The higher the value in the cell in the Markov maps, the darker grey color of each cell become.

Table 1. the features of latent classes (LSMC)

latent class	z_0	z_6
mixture rate	0.13	0.11
purchase rate	2.5%	14.5%

Regarding the latent class z_0, Table 2 shows that the latent class z_0 has high probabilities on self-transition, such as a transition from item to item and that from and category to category compared to the latent class z_6. In the latent class z_6, the probability from top to category is relatively high, while the probability from the top is almost even in the latent class z_0. Therefore, it can be said that the sessions belonging to the latent class z_0 represent the browsing type of “just looking” and those belonging to the latent class z_6 seems to be “willing to purchase”. Considering these results, it is ascertained that the LSMC possibly consider the heterogeneity of the browsing behavior in our targeted EC site.

Table 2. the transition probability of z_0

z_0	blank	cart	cart form	category	conversion	item	login	registration	registration form	sale	search	top
blank	69%	1%	5%	6%	0%	4%	5%	0%	1%	1%	0%	7%
cart	2%	73%	9%	0%	0%	3%	8%	0%	0%	0%	0%	4%
cart form	16%	4%	68%	0%	9%	0%	0%	0%	0%	0%	0%	2%
category	6%	1%	0%	64%	0%	18%	1%	0%	0%	2%	3%	4%
conversion	19%	5%	0%	5%	0%	5%	11%	0%	1%	0%	0%	55%
item	1%	1%	0%	2%	0%	95%	0%	0%	0%	0%	1%	1%
login	12%	2%	4%	2%	0%	1%	53%	0%	4%	0%	0%	19%
registration	4%	9%	1%	6%	0%	1%	7%	15%	9%	0%	0%	48%
resigtration form	3%	0%	1%	0%	0%	0%	7%	6%	77%	0%	0%	5%
sale	3%	2%	0%	13%	0%	15%	1%	0%	0%	59%	3%	3%
search	2%	2%	0%	5%	0%	15%	1%	0%	0%	1%	67%	7%
top	17%	8%	0%	15%	0%	9%	24%	0%	1%	3%	6%	17%

Table 3. the transition probability of z_6

z_6	blank	cart	cart form	category	conversion	item	login	registration	registration form	sale	search	top
blank	26%	10%	3%	13%	0%	40%	2%	0%	0%	1%	1%	5%
cart	6%	26%	6%	4%	0%	42%	10%	0%	0%	0%	0%	6%
cart form	4%	4%	74%	0%	10%	3%	1%	0%	0%	0%	0%	3%
category	2%	1%	0%	55%	0%	37%	0%	0%	0%	1%	1%	2%
conversion	14%	1%	0%	8%	1%	22%	5%	0%	0%	0%	1%	48%
item	5%	20%	0%	31%	0%	37%	0%	0%	0%	1%	2%	4%
login	4%	5%	29%	3%	0%	5%	38%	0%	8%	0%	0%	6%
registration	5%	12%	0%	5%	0%	5%	1%	10%	5%	0%	1%	55%
resigtration form	1%	1%	10%	0%	0%	1%	4%	2%	80%	0%	0%	1%
sale	1%	2%	0%	13%	0%	16%	0%	0%	0%	61%	2%	4%
search	2%	2%	0%	12%	0%	36%	0%	0%	0%	1%	40%	7%
top	5%	7%	0%	50%	0%	20%	4%	0%	0%	3%	5%	6%

4. THE PROPOSED MODEL

4.1. Overview

From the preliminary analysis by the LSMC using the browsing records of the EC site, it is revealed that each latent class successfully considers the heterogeneity of customers' browsing behavior. Moreover, it is found out that the browsing behavior has strong relation to the purchase because the purchase rates are different among latent classes even though the LSMC does not take account of the data of purchase actions. Therefore, by constructing the model that can consider the data of purchase action in addition to the browsing records, it is possible to describe customers' behavior related to the purchase action and to construct better customer segments.

On the other hand, many EC sites conduct various kinds of promotion activities and these measures mainly aim to increase the sales amount. In other words, it is important to analyze that how much purchase rate can be improved by designed promotion activities in order to construct a better promotion strategy. Therefore, if the performance of the promotion activities, such that what kind of browsing behaviors are affected by the activities, is revealed, it is possible to verify the effectiveness of the designed measure and it helps next decision making.

Considering this importance of purchase records, we construct the new latent class model that can take account of purchase records in addition to browsing records based on the LSMC.

4.2. Formulation of the Proposed Model

In this paper, we propose a new latent class model that can take account of the purchase records in addition to the browsing records by extending the LSMC. As with the notations of the LSMC, a set of sessions and page types are defined. Then, a session s_i which has T_i length series of page types can be expressed as $\boldsymbol{x}_i = \left(x_0^i, x_1^i, \cdots, x_{T_i}^i\right)$. Furthermore, we introduce new variable, w_i that indicates the purchase in the i th session as following equation:

$$w_i = \begin{cases} 1, & \text{if an item purchased in } s_i; \\ 0, & otherwise. \end{cases} \tag{4}$$

Here, let the unobserved latent variable be $v_i \in \mathcal{Z} = \{z_l : 1 \le l \le L\}$ that corresponds to the browsing record and the purchase record of the session, s_i, the i th complete data is defined by $(\boldsymbol{x}_i, w_i, v_i)$ and the probability that data $(\boldsymbol{x}_i, w_i)$ is observed is expressed by

$$\begin{aligned} P(\boldsymbol{x}_i, w_i) &= \sum_{v_i \in Z} P(\boldsymbol{x}_i, w_i, v_i) \\ &= \sum_{v_i \in Z} P(v_i) P\left(x_0^i \middle| v_i\right) P(\boldsymbol{x}_i | v_i) \\ &\qquad \cdot P(c|v_i)^{w_i} \mathrm{P}(\bar{c}|v_i)^{1-w_i}, \end{aligned} \tag{5}$$

where c is an event that "an item purchased in s_i" and $\bar{c}$ is complementary event of c. Consequently, assuming first order Markov property to the browsing records, $\boldsymbol{x}_i$, the Equation (5) is extended as

$$\begin{aligned} P(\boldsymbol{x}_i, w_i) &= \sum_{v_i \in Z} P(\boldsymbol{x}_i, w_i, v_i) \\ &= \sum_{v_i \in Z} P(v_i) P\left(x_0^i \middle| v_i\right) \left\{ \prod_{t=1}^{T_i} P\left(x_t^i \middle| x_{t-1}^i, v_i\right) \right\} \\ &\qquad \cdot P(c|v_i)^{w_i} \mathrm{P}(\bar{c}|v_i)^{1-w_i} \end{aligned} \tag{6}$$

Moreover, the occurrence probability of the observed data point, $(\boldsymbol{x}_i, w_i)$ which belongs to a latent class, $v_i = z_l$ is defined by following equation:

$$\begin{aligned} &P(\boldsymbol{x}_i, w_i | v_i = z_l) \\ &= P(\boldsymbol{x}_i | v_i) P(c|v_i)^{w_i} P(\bar{c}|v_i)^{1-w_i} \\ &= \prod_{\mathrm{t}=1}^{\mathrm{T_i}} \lambda_{lj}^{\delta\left(x_0^i = k_j\right)} \prod_{j=1}^{N} \prod_{m=1}^{K} a_{ljm}^{n_{ijm}} \gamma_l^{w_i} (1-\gamma_l)^{1-w_i} \end{aligned} \tag{7}$$

where, $\lambda_{\mathrm{lj}} = P(x_0^i = k_j | v_i = z_l)$, $a_{ljm} = P(x_t = k_j | x_{t-1} = k_m, z_l)$, $\gamma_l = P(c|v_i, z_l)$, n_{ijm} is the i th number of transitions from the page type k_j to the page type k_m, and $\delta(x_0^i = k_j)$ is following indicator function:

$$\delta(x_0^i = k_j) = \begin{cases} 1, & if\ x_0^i = k_j; \\ 0, & otherwise. \end{cases} \tag{8}$$

The graphical model of the proposed model is shown in the Figure 3.

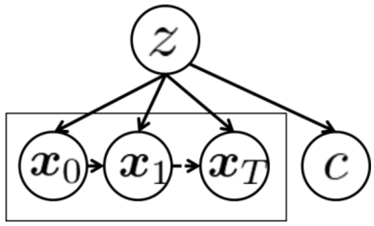

Figure 3. The graphical model of the proposed model

4.3. Parameters Estimation

The proposed model includes unobserved variables, so that the parameters, π_l, λ_l, a_{ljm}, and γ_l are estimated by the expectation maximization (EM) algorithm (Dempster, 1977; McLachlan and Krishnan 2007). The EM algorithm has two steps that are the expectation step (E-step) and the maximization step (M-step). In the E-step, it calculates the expectation value of each unobserved variable on condition that the observed data are given. In the M-step, it updates the parameters using the expectation value of each unobserved variable by update formulas. The parameters of proposed model are estimated by the following formulas of the EM algorithm:

< E-step >

$$P(z_l|\boldsymbol{x}_i, w_i) = \frac{\pi_l P(\boldsymbol{x}_i, w_i|z_l)}{\sum_r^L \pi_r P(\boldsymbol{x}_i, w_i|z_r)} \\ = \alpha_{il} \tag{9}$$

< M-step >

$$\pi_l = \frac{1}{n}\sum_i^I \alpha_{il} \tag{10}$$

$$\lambda_{lj} = \frac{\sum_i^I \alpha_{il}\delta(x_0^i = k_j)}{\sum_i^I \alpha_{il}} \tag{11}$$

$$a_{ljm} = \frac{\sum_i^I \alpha_{il} n_{ijm}}{\sum_m^K \sum_i^I \alpha_{il}\, n_{ijm}} \tag{12}$$

$$\gamma_l = \frac{1}{n\pi_l}\sum_i^I \alpha_{il} w_i \tag{13}$$

The detailed derivations of these formulas are given in Appendix. The EM algorithm repeats E-step and M-step until the log likelihood (the Equation (14)) converges.

$$\log P(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{V}) = \sum_i \log P(\boldsymbol{x}_i, w_i, v_i) \tag{14}$$

In the Equation (14), $\boldsymbol{X}$, $\boldsymbol{W}$, and $\boldsymbol{V}$ indicate $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_I)$, $\boldsymbol{W} = (w_1, w_2, \cdots, w_I)$, and $\boldsymbol{V} = (v_1, v_2, \cdots, v_I)$ respectively.

5. EXPERIMENTS

5.1. Experimental Condition

To examining the effectiveness of our proposed model, we conducted experiments using the same access log data of the Japanese major EC site shown in the Section 3. As one of the promotion activities, the EC site issues discount coupons in real time for customers who are visiting the site. In addition, in order to verify the performance of the discount coupon, the company who managing the EC site conducts A/B testing that divides the coupon target sessions into the "A" group to which is the group coupons are issued and the "B" group to which is the group coupons are not issued, and the difference of "purchase rate" between the groups A and B are obtained by the rate of sessions in which some items are purchased out of coupon in issued/not issued sessions. Combining this A/B testing data and the proposed model, we verify the effectiveness of the proposed model. In order to measure the performance of coupons, we calculate the "performance (A/B)" by the division of purchase rate (A) by purchase rate (B) in each latent class. Moreover, we investigate the parameter, a_{ljm}which indicates the transition probabilities from the page type k_j to k_m for characterizing each latent class in addition to mixture rate (π_l) and purchase probability (c_l). From the results of preliminary experiments, the number of latent classes is set as 8.

5.2. Results and Discussion

Focusing on the parameter $P(c|z)$ which indicates the purchase probability, and the differences between the "A" group and the "B" group, 3 latent classes out of 8 are selected in the Table 4.

Table 4. the features of latent classes

Latent class	z_2	z_4	z_6
Mixture rate	0.06	0.08	0.27
Purchase probability (c_l)	0.092	0.481	0.002
Coupon target sessions rate	3.3%	6.2%	4.0%
Purchase rate (A)	25.2%	68.8%	0.8%
purchase rate (B)	16.7%	63.3%	0.9%
performance (A/B)	1.51	1.09	0.89

Table 4 shows that the largest number of coupons are issued to the latent class z_4. In this latent class z_4, the "purchase probability" shows a high value such as 0.481 and the "performance (A/B)", 1.09 is not so high; therefore, the discount coupon for this latent class is not very effective. In other words, sessions belonging to this latent class contains cases which purchase some items without issuing coupons, so it is not reasonable to issue coupons on a session belonging to this latent class. On the other hand, in the latent class z_2, the "purchase probability" shows relatively low value, 0.092 and the "performance (A/B)", 1.51 is the highest among all latent classes. This result implies that the sessions belonging to this latent class z_2 indicate positive response for discount coupons. Therefore, it can conclude that issuing discount coupons should be focused to the latent class z_2 than the latent class z_4. Furthermore, comparing Markov maps of z_2 and z_4 shown in Table 5 and 6, most transition probabilities of z_4 are similar with z_2 but transition probabilities which are related to purchase such as from "item" to "cart", from "login" to "cart form" and from "cart form" to "conversion" are higher than z_2. This results indicate that browsing behavior of z_2 and z_4 are almost similar except purchase. In other words, sessions belonging to z_2 are about to purchase but they do not. Therefore, issuing discount coupons to z_2 is effective and this result can be considered as one of the reasons why "performance (A/B)" of z_2 is high.

Concerning the latent class z_6, both of the "purchase rate" and the "performance (A/B)" show low value. Especially, the value of "performance (A/B)" is less than 1 means that issuing discount coupon diminishes the purchase rate. Therefore, it is not reasonable to issue coupons to the latent class z_6.

Table 5. the parameter $a_{ljm}$of z_2

z_2	blank	cart	cart form	category	conversion	item	login	registration	resistration form	sale	search	top
blank	65%	3%	6%	7%	0%	9%	2%	0%	0%	4%	0%	4%
cart	2%	42%	5%	5%	0%	32%	5%	0%	0%	3%	0%	6%
cart form	15%	3%	62%	0%	15%	1%	1%	0%	0%	0%	0%	2%
category	1%	1%	0%	57%	0%	34%	0%	0%	0%	4%	1%	2%
conversion	11%	1%	0%	6%	1%	7%	9%	0%	0%	2%	0%	62%
item	2%	11%	0%	40%	0%	7%	0%	0%	0%	36%	2%	1%
login	10%	3%	11%	3%	0%	2%	45%	0%	2%	2%	0%	22%
registration	6%	13%	0%	5%	0%	0%	0%	9%	9%	0%	0%	58%
resistration form	2%	0%	3%	0%	0%	0%	6%	7%	76%	0%	0%	5%
sale	1%	1%	0%	5%	0%	35%	0%	0%	0%	58%	1%	1%
search	1%	2%	0%	11%	0%	25%	0%	0%	0%	7%	49%	5%
top	7%	5%	0%	39%	0%	4%	16%	0%	0%	18%	3%	7%

Table 6. the parameter $a_{ljm}$of z_6

z_4	blank	cart	cart form	category	conversion	item	login	registration	registration form	sale	search	top
blank	26%	13%	7%	15%	1%	27%	4%	0%	2%	1%	1%	4%
cart	3%	24%	6%	7%	0%	43%	12%	0%	0%	0%	0%	5%
cart form	4%	3%	65%	0%	23%	2%	1%	0%	0%	0%	0%	2%
category	1%	2%	0%	54%	0%	40%	0%	0%	0%	0%	1%	1%
conversion	10%	1%	0%	8%	1%	13%	8%	0%	0%	0%	0%	57%
item	2%	30%	0%	55%	0%	8%	0%	0%	0%	1%	2%	1%
login	3%	4%	33%	3%	0%	2%	41%	0%	7%	0%	0%	5%
registration	2%	14%	1%	5%	0%	2%	3%	10%	6%	0%	1%	55%
registration form	1%	1%	10%	0%	0%	0%	5%	2%	79%	0%	0%	1%
sale	1%	3%	0%	13%	0%	13%	0%	0%	0%	64%	3%	3%
search	1%	3%	0%	14%	0%	37%	0%	0%	0%	1%	39%	4%
top	4%	11%	0%	59%	0%	7%	6%	0%	0%	2%	4%	6%

6. CONCLUSION AND FUTURE WORK

We proposed a new latent class model that takes account of the time series of browsing records and the purchase records. In addition, we showed the effectiveness of the proposed model by demonstrating the data analysis which brought out effective next actions for issuing better discount coupons to appropriate target using the access log data of a Japanese major EC site. For examining the effectiveness of the proposed model, we investigated one promotion activity, that is, the real-timing discount coupon of one EC site. Conducting experiments for different kind of activities on different EC sites are our future work.

Furthermore, we constructed a model extending the LSMC which is based on the simplest latent class model. To improve the analysis, it is possible to construct more complex types of latent class models such as the Flexible Mixture Model (Si and Jin, 2003) which assumes two different types of latent variable, the Latent Dirichlet Allocation (Blei et al., 2003) which uses a hyper-parameter for the latent class. This task also comprises our future work.

ACKNOWLEDGMENTS

The authors would like to express our thanks to Dr. Haruka Yamashita, and all members of Goto Laboratory, Waseda University, for their helpful comments regarding this research. In addition, we wish to thank offering the access log data of the EC site and timely kind help given by Emotion Intelligence, Inc. A part of this study was supported by JSPS KAKENHI Grant Numbers 26282090 and 26560167.

APPENDIX

The Equations (9) – (13), which are needed to apply the EM algorithm to the proposed model, are derived in detail here.

< E-step >

In the E-step, the probability, $\tilde{P}(\boldsymbol{V}|\boldsymbol{X},\boldsymbol{W})$ that is the posterior probability of the latent variable can be obtained by the following equation:

$$\begin{aligned}\tilde{P}(\boldsymbol{V}|\boldsymbol{X},\boldsymbol{W}) &= \frac{\tilde{P}(\boldsymbol{X},\boldsymbol{W},\boldsymbol{V})}{\tilde{P}(\boldsymbol{X},\boldsymbol{W})} \\ &= \prod_{i=1}^{I} \frac{\pi_l P(\boldsymbol{x}_i, w_i|z_l)}{\sum_{r=1}^{L} \pi_r P(\boldsymbol{x}_i, w_i|z_r)} \\ &= \prod_{i=1}^{I} \alpha_{il}. \end{aligned} \tag{15}$$

< M-step >

Using α_{il} derived in the E-step, maximize the Q function that is lower limit of log-likelihood, $\log P(\boldsymbol{X},\boldsymbol{W},\boldsymbol{V})$ in the M-step with respect to parameters, π_l, λ_l, a_{ljm}, and γ_l.

The Q function is defined by the following equation:

$$Q = \sum_{\boldsymbol{V}} P(\boldsymbol{V}|\boldsymbol{X},\boldsymbol{W}) \log P(\boldsymbol{V}|\boldsymbol{X},\boldsymbol{W}), \tag{16}$$

and can be expanded as follows:

$$\begin{aligned} Q &= \sum_{i=1}^{I}\sum_{l=1}^{L} \alpha_{il} \log \pi_l + \sum_{i=1}^{I}\sum_{l=1}^{L} \alpha_{il} P(\boldsymbol{x}_i, w_i|z_l) \\ &= \sum_{i=1}^{I}\sum_{l=1}^{L} \alpha_{il} \log \pi_l \end{aligned}$$

$$+\sum_{i=1}^{I}\sum_{l=1}^{L}\alpha_{il}\sum_{j=1}^{J}\delta(x_0^i = k_j)\log\lambda_{lj}$$

$$+\sum_{i=1}^{I}\sum_{l=1}^{L}\alpha_{il}\sum_{j=1}^{J}\sum_{m=1}^{J} n_{ijm}\log a_{ljm}$$

$$+\sum_{i=1}^{I}\sum_{l=1}^{L}\alpha_{il}w_i\log\gamma_l$$

$$+\sum_{i=1}^{I}\sum_{l=1}^{L}\alpha_{il}(1-w_i)\log(1-\gamma_l). \tag{17}$$

Here, the constraints of the parameters are given by:

$$\sum_{l=1}^{L}\pi_l = 1, \tag{18}$$

$$\sum_{i=1}^{J}\lambda_{lj} = 1 \quad \text{for } \forall z_l \in \mathcal{Z}, \tag{19}$$

$$\sum_{m=1}^{J} a_{ljm} = 1 \quad \text{for } \forall z_l \in \mathcal{Z} \text{ and } \forall k_j \in \mathcal{K}, \tag{20}$$

$$\gamma_l + \overline{\gamma_l} = 1 \quad \text{for } \forall z_l \in \mathcal{Z}. \tag{21}$$

Applying the Lagrange multiplier method, the objective function LM is obtained as

$$LM = Q - \eta\left(\sum_{l=1}^{L}\pi_l - 1\right) - \sum_{l=1}^{L}\xi_l\left(\sum_{j=1}^{J}\lambda_{lj} - 1\right)$$
$$-\sum_{l=1}^{L}\sum_{j=1}^{J}\zeta_{lj}\left(\sum_{m=1}^{J}a_{ljm} - 1\right)$$
$$-\sum_{l=1}^{L}\iota_l(\gamma_l + \bar{\gamma}_l - 1). \tag{22}$$

The optimized π_l is derived by differentiating the objective function LM with respect to π_l as follows:

$$\frac{\partial LM}{\partial \pi_l} = \sum_{i=1}^{I}\alpha_{il}\frac{1}{\pi_l} - \eta = 0$$
$$\pi_l = \frac{1}{\eta}\sum_{i=1}^{I}\alpha_{il}\,. \tag{23}$$

Clearly, the following equation

$$\sum_{l=1}^{L}\pi_l = \sum_{l=1}^{L}\frac{1}{\eta}\sum_{i=1}^{I}\alpha_{il} = 1, \tag{24}$$

is satisfied by the Equation (18), where η is

$$\eta = \frac{1}{n}. \tag{25}$$

Considering the Equations (23) and (25), we have

$$\pi_l = \frac{1}{n}\sum_{i=1}^{I}\alpha_{il}\,. \tag{26}$$

The optimized λ_{lj} is derived by differentiating the objective function LM with respect to λ_{lj} as follows:

$$\frac{\partial LM}{\partial \lambda_{lj}} = \sum_{i=1}^{I}\alpha_{il}\delta(x_0^i = k_j)\frac{1}{\lambda_{lj}} - \xi_l = 0$$
$$\lambda_{lj} = \frac{1}{\xi_l}\sum_{i=1}^{I}\alpha_{il}\delta(x_0^i = k_j)\,. \tag{27}$$

Clearly, the following equation

$$\sum_{j=1}^{J}\lambda_{lj} = \sum_{j=1}^{J}\frac{1}{\xi_l}\sum_{i=1}^{I}\alpha_{il}\delta(x_0^i = k_j) = 1, \tag{28}$$

is satisfied by the Equation (19), where ξ_l is

$$\xi_l = \sum_{i=1}^{I}\alpha_{il}\,. \tag{29}$$

Considering the Equations (27) and (29), we have

$$\lambda_{lj} = \frac{\sum_{i=1}^{I}\alpha_{il}\delta(x_0^i = k_j)}{\sum_{i=1}^{I}\alpha_{il}}. \tag{30}$$

The optimized a_{ljm} is derived by differentiating the objective function LM with respect to a_{ljm} as follows:

$$\frac{\partial LM}{\partial a_{ljm}} = \sum_{i=1}^{I}\alpha_{il}n_{ijl}\frac{1}{\zeta_{lj}} - \zeta_{lj} = 0$$
$$a_{ljm} = \frac{1}{\zeta_{lj}}\sum_{i=1}^{I}\alpha_{il}n_{ijl}\,. \tag{31}$$

Clearly, the following equation

$$\sum_{m=1}^{J}a_{ijm} = \sum_{m=1}^{J}\frac{1}{\zeta_{lj}}\sum_{i=1}^{I}\alpha_{il}n_{ijl} = 1, \tag{32}$$

is satisfied by the Equation (20), where ζ_{lj} is

$$\zeta_{lj} = \sum_{m=1}^{J}\sum_{i=1}^{I}\alpha_{il}n_{ijl}\,. \tag{33}$$

Considering the Equations (31) and (33), we have

$$a_{ljm} = \frac{\sum_{i=1}^{I} \alpha_{il} n_{ijl}}{\sum_{m=1}^{J} \sum_{i=1}^{I} \alpha_{il} n_{ijl}}. \tag{34}$$

The optimized c_l is derived by differentiating the objective function LM with respect to c_l as follows:

$$\begin{aligned} \frac{\partial LM}{\partial c_l} &= \sum_{i=1}^{I} \alpha_{il} w_i \frac{1}{c_l} - \iota_l = 0 \\ c_l &= \frac{1}{\iota_l} \sum_{i=1}^{I} \alpha_{il} w_i . \end{aligned} \tag{35}$$

Similarly, for $\bar{c}_l$,

$$\begin{aligned} \frac{\partial LM}{\partial \bar{c}_l} &= \sum_{i=1}^{I} \alpha_{il} (1 - w_i) \frac{1}{\bar{c}_l} - \iota_l = 0 \\ \bar{c}_l &= \frac{1}{\iota_l} \sum_{i=1}^{I} \alpha_{il} (1 - w_i) . \end{aligned} \tag{36}$$

Clearly, the following equation

$$\frac{1}{\iota_l} \sum_{i=1}^{I} \alpha_{il} w_i \ + \frac{1}{\iota_l} \sum_{i=1}^{I} \alpha_{il} (1 - w_i) = 1 , \tag{37}$$

is satisfied by the Equation (21), where η is

$$\iota_l = \sum_{i=1}^{I} \alpha_{il} = n\pi_l . \tag{38}$$

Considering the Equations (36) and (38), we have

$$c_l = \frac{1}{n\pi_l} \sum_{i=1}^{I} \alpha_{il} w_i , \tag{39}$$

REFERENCES

Burt, S. and Sparks, L. (2003) E-commerce and the retail process: a review, *J. Retailing and Consumer Services*, Vol. 10, 275—286.

Wedel, M. and Kamakura, A. W. (2012) Market segmentation: Conceptual and methodological foundations, Springer.

Plant, T. R. (2000) Ecommerce: Formulation of Strategy, Prentice Hall Professional.

Thorsten, B. and Francine, C. (2002) Topic-based document segmentation with probabilistic latent semantic analysis, *Proc. the eleventh international conf. Information and knowledge management*, 211—218.

Bhatnagar, A. and Ghose, S (2004) A latent class segmentation analysis of e-shoppers, *J. Business Research*, Vol. 57, 758—767.

Bassi, F. (2007) Latent class factor models for market segmentation: An application to pharmaceuticals, *Statis- tical Methods & Applications*, Vol. 16 No. 2, 279—287.

Hofmann, T. (2004) Latent semantic models for collaborative filtering, *J. ACM Trans. Information Systems*, Vol. 22 Issue 1, 89—115.

Yamagami, K. Fujiwara, N. Mikawa , K. and Goto, M. (2014) A Statistical Model for Recommender System to Maximize Sales Amount Focusing on Characteristics of EC Site Data, *Proc. 15th Asia Pacific Industrial Engineering and Management Systems Conference*

Matsuzaki, Y. Yamagami, K, Mikawa, K and Goto, M. (2010) Analysis of Customer Purchase Behavior by using Purchase History with Discount Coupon Based on Latent Class Model, *Proc. 16th Asia Pacific Industrial Engineering and Management Systems Conference*.

Dias, J. G. and Vermunt, J. K. (2007) Latent class modeling of website users' search patterns: Implications for online market segmentation, *J. Retailing and Consumer Services 14*, 359—368.

Hofmann, T. (1999) Probabilistic Latent Semantic Indexing, *Proc. the 22nd Annual International SIGIR Conf. Research and Development in Information Retrieval*, 50—57.

Mei , Q. and Zhai, C. (2006) A Mixture Model for Contextual Text Mining, *Proc. 12th ACM SIGKDD International Conf. Knowledge Discovery and Data Mining*, 649—655.

Ishigaki, T. Takenaka , T. and Motomura, Y. (2010) Category Mining by Heterogeneous Data Fusion Using PdLSI Model in a Retail Service, *Proc. IEEE Inter. Conf. Data Mining*, 857—862.

Hofmann, T. (2003) Collaborative filtering via gaussian probabilistic latent semantic analysis, *Proc. the 26th annual international ACM SIR conf. Research and development in information retrieval*, 259—266.

Jin, R., Si, L. and Zhai, C. (2003) Preference-based Graphic Models for Collaborative Filtering, *Proc. the 19th Conf. Uncertainty in Artificial Intelligence*, 329—336.

Jin, R., Si, L. and Zhai, C. (2006) A Study of Mixture Models for Collaborative Filtering, *J. Information Retrieval*, Vol. 9, No. 3, 357—382.

Si , L. and Jin, R. (2003) Flexible Mixture Model for Collaborative Filtering, *Proc. 20th International Conf. Machine Learning*, Volume 3, 704—711.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003) Latent Dirichlet allocation, *J. Machine Learning Research*, Vol. 3, 993—102.