A Modified Aspect Model for Simulation Analysis

Masayuki Goto*, Kazushi Minetoma*, Kenta Mikawa*, Manabu Kobayashi†, and Shigeichi Hirasawa‡
*School of Creative Science and Engineering, Waseda University, Tokyo, JAPAN
Email: goto@it.mgmt.waseda.jp
†Department of Information Management Science,
Shonan Institute of Technology, Kanagawa, JAPAN
‡Research Institute for Science and Engineering, Waseda University, Tokyo, JAPAN

***Abstract*—This paper proposes a new latent class model to represent user segments in a marketing model of electric commerce sites. The aspect model proposed by T. Hofmann is well known and is also called the probabilistic latent semantic indexing (PLSI) model. Although the aspect model is one of effective models for information retrieval, it is difficult to interpret the meaning of the probability of latent class in terms of marketing models. It is desirable that the probability of latent class means the size of customer segment for the purpose of marketing research. Through this formulation, the simulation analysis to dissect the several situations become possible by using the estimated model. The impact of the strategy that we contact to the specific customer segment and make effort to increase the number of customers belonging to this segment can be predicted by using the model demonstrating the size of customer segment. This paper proposes a new model whose probability parameter of latent variable means the rate of users with the same preference in market. By applying the proposed model to the data of an internet portal site for job hunting, the effectiveness of our proposal is verified.**

I. Introduction

Probabilistic Latent Semantic Analysis (PLSA) is an effective and interesting approach to automated document indexing and information retrieval [1]. The PLSA is beased on a statistical model which is called the *aspect model*. The aspect model is also one of latent class models which are oftentimes applied in the field of marketing [2]-[4]. The aspect model is a statistical mixture model with the latent variable taking a discrete hidden class behind the observed pair data. Although the aspect model was applied to analyze the text data, the application is not restricted into automated document indexing and information retrieval. This model is effective for the purpose of recommendation [5],[6] because it is enough to acquire the precise prediction of users' rating or preference by using the aspect model [7].

On the other hand, the interpretation of the latent variable is important when it is regarded as the market model [8]. This is because an effective marketing strategy is sometimes discussed by using the estimated model. However, the probability of the latent class of the aspect model cannot be interpreted clearly in the marketing sense. The probability of the latent class of the aspect model does not mean the size of a segment of similar consumers in a market. In terms of the marketing model, it is desirable that the probability of latent class means the size of customer segment. Moreover, the good customers cannot be discriminated in the aspect model. Usually, there are good customers who purchase many items in a market and it is the important target in the marketing purpose. However, even if the frequency levels of purchase action are quite different between some users, the users may belong to the same latent class when they bought the similar items. From the viewpoint of marketing analysis, it is desirable to discriminate the good customers from others.

This paper proposes a new model whose probability parameter of latent variable means the segment size, that is, the rate of users with the same preference in market. Through the formulation proposed in this paper, the simulation analysis to dissect the several situations become possible by using the estimated model. For example, the impact of the strategy that we contact to the specific customer segment and make effort to increase the number of customers belonging to this segment can be predicted by using the model demonstrating the size of customer segment. By applying the proposed model to the data of an internet portal site for job hunting, the effectiveness of our proposal is verified.

II. Probabilistic Latent Semantic Indexing

A. Aspect Model

The aspect model [1] is a latent variable model for general co-occurrence data which associates an unobserved class variable $z \in \mathcal{Z} = \{z_1, z_2, \cdots, z_Z\}$ with each observation. The aspect model was introduced in Probabilistic Latent Semantic Analysis to automated document indexing and information retrieval. The aspect model is a statistical mixture model to represent the relation between a word $w \in \mathcal{W} = \{w_1, w_2, \cdots, w_W\}$ and a document $d \in \mathcal{D} = \{d_1, d_2, \cdots, d_D\}$. The generative model of the aspect model is defined in the following way [1]:

1) select a document d with probability $P(d)$,
2) pick a latent class z with probability $P(z|d)$,
3) generate a word w with probability $P(w|z)$.

The latent class variable cannot be observed by analysts but its occurrence is derived following the probability $P(z|d)$. That is, an unobserved class z is latently defined for each observed pair (d, w). The generative model of event by this way is expressed by

$$P(d, w) = P(d) \sum_{z \in \mathcal{Z}} P(w|z)P(z|d). \tag{1}$$

The equivalent symmetric version of the model is

$$P(d, w) = \sum_{z \in \mathcal{Z}} P(w|z)P(d|z)P(z), \tag{2}$$

which is a re-parameterized version of the generative model. The model in the form Eq.(2) is convenient for parameter estimation by EM algorithm.

B. Model Learning by EM Algorithm

The maximum likelihood estimation in latent variable models can be calculated by the Expectation Maximization (EM) algorithm [9],[10] The log likelihood function is given by

$$L = \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} n(d,w) \log P(d,w), \tag{3}$$

where $n(d,w)$ denotes the term frequency defined by the number of times w occurred in d in the training data set.

The EM algorithm for the aspect model is given as follows:

[E-Step]

$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(d|z')P(w|z')}. \tag{4}$$

[M-Step]

$$P(w|z) = \frac{\sum_{d'} n(d',w)P(z|d',w)}{\sum_{d'}\sum_{w'} n(d',w')P(z|d',w')}. \tag{5}$$

$$P(d|z) = \frac{\sum_{w'} n(d,w')P(z|d,w')}{\sum_{d'}\sum_{w'} n(d',w')P(z|d',w')}. \tag{6}$$

$$P(z) = \frac{1}{\sum_{d'}\sum_{w'} n(d',w')} \sum_{d'} \sum_{w'} n(d',w')P(z|d',w'). \tag{7}$$

C. Aspect Model for Web Marketing

The aspect model can be applied to the purpose of recommendation which is one of the effective web marketing tools [7]. The recommender system is an information filtering system to predict *rating* or *preference* that each user would evaluate to each item. Recently, the recommender system has become one of the strong web marketing tools The aspect model gives a way of model-based collaborative filtering. D. Chen et. al. [11] applied the aspect model to cluster user's web pages and construct user profile for implementing personalized recommendation.

Here, let the set of users be $\mathcal{U} = \{u_1, u_2, \cdots, u_U\}$, the set of product items be $\mathcal{A} = \{a_1, a_2, \cdots, a_A\}$. As a model of the purchase behavior of users, the aspect model gives the following expression:

$$P(u,a) = \sum_{z \in \mathcal{Z}} P(u|z)P(a|z)P(z), \tag{8}$$

which means the probability such that the user u purchase the item a. This model is effective for the purpose of recommendation because it is enough to acquire the precise prediction of users' rating or preference by using the aspect model. Sometimes, the users can be clustered in terms of preference to product items [8]. The aspect model can be useful to represent the situation of the market model with latent variables. In the research field of marketing science, the behavioral process of customers' product choice has been studied and various kinds of mixed model, e.g. the mixed logit model, the mixed nested logit model, the mixed probit model, etc. were proposed [8].

On the other hand, the interpretation of the latent variable is important when it is regarded as the market model. This is because an effective marketing strategy is sometimes discussed by using the estimated model. However, the probability of the latent variable, $P(z)$, is not clearly interpreted in the marketing sense. $P(z)$ does not mean the size of a segmentation of similar consumers in a market.

The probability $P(a|z)$ is the parameter of the multinomial distribution conditioned by the latent class z and it means the probability that the item a is purchased by a user belonging to the latent class z. The probability $P(u|z)$ is the probability that the user u purchase some item under the condition that the latent class z is fixed. Usually, there are good customers who purchase many items in a market and it is the important target in the marketing purpose. The frequency information of purchase by each user is taken into the probability $P(u|z)$ in the learning phase. Even if the frequency levels of purchase action are quite different between some users, the users may belong to the same latent class when they bought the similar items. The good customers cannot be discriminated in the aspect model.

III. The Proposed Model

As mentioned in the previous section, the aspect model has several limitations for the purpose of marketing model.

1) The good customers cannot be discriminated from other customers, although the good customers are important targets in the marketing sense.
2) The probability of the latent variable is not clearly interpreted in the marketing sense.

Because of these problems, it is difficult to conduct the simulation analysis with strategic consideration by using the estimated aspect model. Therefore, we propose a new modified version of the aspect model to overcome these problems.

A. Formulation of Model

The purchase history of a user u is denoted by the following vector:

$$\boldsymbol{x}_u = (x_1^u, x_2^u, x_3^u, \cdots, x_A^u), \tag{9}$$

where

$$x_j^u = \begin{cases} 1, & \text{when the item} a_j \text{has been purchased by the user } u \\ 0, & \text{when the item} a_j \text{has not been purchased by the user } u \text{yet} \end{cases} \tag{10}$$

The probability $P(z)$ is the parameter of the multinomial distribution on the discrete set $\mathcal{Z}$ and $P(u|z)$ is that on the set $\mathcal{U}$. On the other hand, we define the probability $P(a|z)$ as the binomial probability for the binary events:

1) a: The item a is purchased.
2) $\bar{a}$: The item a is not purchased.

That is, $P(a|z) + P(\bar{a}|z) = 1$ is satisfied.

The probability model is defined by

$$P(u, \boldsymbol{x}_u, z) = P(z)P(u|z) \prod_{j=1}^{A} P(a|z)^{x_j^u} P(\bar{a}|z)^{1-x_j^u}, \tag{11}$$

where $u \in \mathcal{U}, \;\; a \in \mathcal{A}, \;\; z \in \mathcal{Z}$.

On this settings, the complete probability model of purchase action for all U users is given by

$$P(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{V}) = \prod_{k=1}^{U} P(v_k)P(u_k|v_k) \prod_{j=1}^{A} P(a_j|v_k)^{x_j^{u_k}} P(\bar{a}_j|v_k)^{1-x_j^{u_k}}, \tag{12}$$

where v_k is the latent class of the user u_k, and $v_k \in \mathcal{Z}$, $\boldsymbol{Y} = (u_1, u_2, \cdots, u_U)^T$, $\boldsymbol{X} = (\boldsymbol{x}_{u_1}, \boldsymbol{x}_{u_2}, \cdots, \boldsymbol{x}_{u_U})$, $\boldsymbol{V} = (v_1, v_2, \cdots, v_U)^T$.

B. Construction of EM algorithm

Because Eq.(11) includes the unobserved latent class, the maximum likelihood estimator cannot be formulated by a clear equation. In order to estimate the parameters $P(z)$, $P(u|z)$, and $P(\boldsymbol{x}|z)$, the EM algorithm is applied.

1) E-Step: The probability $\tilde{P}(\boldsymbol{V}|\boldsymbol{Y}, \boldsymbol{X})$ which is necessary to calculate the expectation in E-step should be calculated.

$$\begin{aligned}
&\tilde{P}(\boldsymbol{V}|\boldsymbol{Y}, \boldsymbol{X}) \\
&= \frac{\prod_{k=1}^{U} \tilde{P}(v_k)\tilde{P}(u_k|v_k) \prod_{j=1}^{A} \tilde{P}(a_j|v_k)^{x_j^{u_k}} \tilde{P}(\bar{a}_j|v_k)^{1-x_j^{u_k}}}{\sum_{v_k \in \mathcal{Z}} \prod_{k=1}^{U} \tilde{P}(v_k)\tilde{P}(u_k|v_k) \prod_{j=1}^{A} \tilde{P}(a_j|v_k)^{x_j^{u_k}} \tilde{P}(\bar{a}_j|v_k)^{1-x_j^{u_k}}} \\
&= \prod_{k=1}^{U} \frac{\tilde{P}(v_k)\tilde{P}(u_k|v_k) \prod_{j=1}^{A} \tilde{P}(a_j|v_k)^{x_j^{u_k}} \tilde{P}(\bar{a}_j|v_k)^{1-x_j^{u_k}}}{\sum_{v_k \in \mathcal{Z}} \tilde{P}(v_k)\tilde{P}(u_k|v_k) \prod_{j=1}^{A} \tilde{P}(a_j|v_k)^{x_j^{u_k}} \tilde{P}(\bar{a}_j|v_k)^{1-x_j^{u_k}}} \\
&= \prod_{k=1}^{U} \tilde{P}(v_k|u_k, \boldsymbol{x}_u). \qquad (13)
\end{aligned}$$

Then, the Q-function

$$Q = \sum_{\boldsymbol{V}} \tilde{P}(\boldsymbol{V}|\boldsymbol{Y}, \boldsymbol{X}) \log P(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{V}), \tag{14}$$

is expanded as follows.

$$\begin{aligned}
Q &= \sum_{\boldsymbol{V}} \tilde{P}(\boldsymbol{V}|\boldsymbol{Y}, \boldsymbol{X}) \\
&\quad \cdot \log \left\{ \prod_{k=1}^{U} P(v_k)P(u_k|v_k) \prod_{j=1}^{A} P(a_j|v_k)^{x_j^{u_k}} P(\bar{a}_j|v_k)^{1-x_j^{u_k}} \right\} \\
&= \sum_{\boldsymbol{V}} \tilde{P}(\boldsymbol{V}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}) \sum_{k=1}^{U} \Big\{ \log P(v_k) + \log P(u_k|v_k) \\
&\qquad + \sum_{j=1}^{A} \Big(x_j^{u_k} \log P(a_j|v_k) + (1 - x_j^{u_k}) \log P(\bar{a}_j|v_k) \Big) \Big\} \\
&= \sum_{k=1}^{U} \sum_{z \in \mathcal{Z}} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) \log P(z) \\
&\quad + \sum_{k=1}^{U} \sum_{z \in \mathcal{Z}} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) \log P(u_k|z) \\
&\quad + \sum_{k=1}^{U} \sum_{z \in \mathcal{Z}} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) \\
&\qquad \cdot \sum_{j=1}^{A} \Big(x_j^{u_k} \log P(a_j|z) + (1 - x_j^{u_k}) \log P(\bar{a}_j|z) \Big) \\
&= \sum_{k=1}^{U} \sum_{z \in \mathcal{Z}} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) \log P(z) \\
&\quad + \sum_{k=1}^{U} \sum_{z \in \mathcal{Z}} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) \log P(u_k|z) \\
&\quad + \sum_{z \in \mathcal{Z}} \sum_{j=1}^{A} \sum_{k=1}^{U} \Big(\tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) x_j^{u_k} \log P(a_j|z) \\
&\qquad + (1 - x_j^{u_k}) \log P(\bar{a}_j|z) \Big). \qquad (15)
\end{aligned}$$

2) M-Step: The constraints of the parameters are given by

$$\begin{aligned}
\sum_{z \in \mathcal{Z}} P(z) &= 1, & (16) \\
\sum_{u \in \mathcal{U}} P(u|z) &= 1 \quad \text{for} \quad \forall z \in \mathcal{Z}, & (17) \\
P(a|z) + P(\bar{a}|z) &= 1 \quad \text{for} \quad \forall z \in \mathcal{Z},\ \forall a \in \mathcal{A}, & (18) \\
\sum_{\boldsymbol{w} \in \mathcal{W}} P(\boldsymbol{w}|z) &= 1 \quad \text{for} \quad \forall z \in \mathcal{Z}. & (19)
\end{aligned}$$

Applying the Lagrange multipliers method, the objective function is derived as

$$\begin{aligned}
L &= Q - \alpha \left(\sum_{z \in \mathcal{Z}} P(z) - 1 \right) - \sum_{z \in \mathcal{Z}} \beta_z \left(\sum_{u \in \mathcal{U}} P(u|z) - 1 \right) \\
&\quad - \sum_{a \in \mathcal{A}} \sum_{z \in \mathcal{Z}} \gamma_{za} (P(a|z) + P(\bar{a}|z) - 1) \\
&\quad - \sum_{z \in \mathcal{Z}} \lambda_z \left(P(\boldsymbol{w}|z) - 1 \right). \qquad (20)
\end{aligned}$$

(1) The optimization of $P(z)$:
By the equation

$$\frac{\partial L}{\partial P(z)} = \sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) \frac{1}{P(z)} - \alpha = 0, \tag{21}$$

we have

$$P(z) = \frac{1}{\alpha} \sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}). \tag{22}$$

Because

$$\sum_{z \in \mathcal{Z}} P(z) = \sum_{z \in \mathcal{Z}} \frac{1}{\alpha} \sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) = 1, \tag{23}$$

is satisfied obviously by Eq.(16), α is calculated by

$$\alpha = \sum_{z \in \mathcal{Z}} \sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) = U. \tag{24}$$

Therefore we have

$$P(z) = \frac{1}{U} \sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}). \tag{25}$$

(2) The optimization of $P(u_k|z)$:
By the equation

$$\frac{\partial L}{\partial P(u_k|z)} = \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) \frac{1}{P(u_k|z)} - \beta_z = 0, \tag{26}$$

we have

$$P(u_k|z) = \frac{1}{\beta_z}\tilde{P}(z|u_k, \boldsymbol{x}_{u_k}). \tag{27}$$

Eq.(17) derives

$$\sum_{k=1}^{U} P(u_k|z) = \sum_{k=1}^{U} \frac{1}{\beta_z}\tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) = 1. \tag{28}$$

Therefore, by the equation

$$\beta_z \quad = \quad \sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}), \tag{29}$$

we have

$$P(u_k|z) = \frac{1}{\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k})}\tilde{P}(z|u_k, \boldsymbol{x}_{u_k}). \tag{30}$$

(3) The optimization of $P(a_j|z)$:
By the equations

$$\begin{aligned}\frac{\partial L}{\partial P(a_j|z)} &= \left(\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) x_j^{u_k}\right) \frac{1}{P(a_j|z)} - \gamma_{za_j} = 0, \\ \frac{\partial L}{\partial P(\bar{a}_j|z)} &= \left(\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) \left(1 - x_j^{u_k}\right)\right) \frac{1}{P(\bar{a}_j|z)} - \gamma_{za_j} = 0,\end{aligned}$$

we have

$$P(a_j|z) \quad = \quad \frac{1}{\gamma_{za_j}} \left(\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) x_j^{u_k}\right), \tag{31}$$

$$P(\bar{a}_j|z) \quad = \quad \frac{1}{\gamma_{za_j}} \left(\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) \left(1 - x_j^{u_k}\right)\right). \tag{32}$$

Eq.(18) means

$$\begin{aligned} P(a_j|z) + P(\bar{a}_j|z) &= \frac{1}{\gamma_{za_j}} \sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) \\ &= 1. \end{aligned} \tag{33}$$

Therefore, we have

$$\gamma_{za_j} \quad = \quad \sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}). \tag{34}$$

Finally, we have

$$P(a_j|z) \quad = \quad \frac{1}{\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k})} \left(\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) x_j^{u_k}\right), \tag{35}$$

$$P(\bar{a}_j|z) \quad = \quad \frac{1}{\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k})} \left(\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) \left(1 - x_j^{u_k}\right)\right). \tag{36}$$

3) EM Algorithm: The EM algorithm to estimate the parameters of the proposed model is given by as follows:

[E-Step]

$$\tilde{P}(z|u_k, \boldsymbol{x}_{u_k}), \tag{37}$$

[M-Step]

$$P(z) = \frac{1}{U}\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}), \tag{38}$$

$$P(u_k|z) = \frac{1}{\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k})}\tilde{P}(z|u_k, \boldsymbol{x}_{u_k}), \tag{39}$$

$$P(a_j|z) \quad = \quad \frac{1}{\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k})} \left(\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) x_j^{u_k}\right), \tag{40}$$

$$P(\bar{a}_j|z) \quad = \quad \frac{1}{\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k})} \left(\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}) \left(1 - x_j^{u_k}\right)\right). \tag{41}$$

IV. Consideration of the Proposed Model

Let us consider the meaning of the probability

$$P(a_j|z) = \frac{1}{\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}, \boldsymbol{s}_{u_k})} \left(\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}, \boldsymbol{s}_{u_k}) x_j^{u_k}\right), \tag{42}$$

of the proposed latent class model.

The denominator $\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}, \boldsymbol{s}_{u_k})$ of this equation is meaning the expected number of users belonging to the latent class z. That is, the probability $P(a_j|z)$ means the rate of users who would purchase the item a_j in all users belonging to the latent class z. Therefore,

$$N_z(z) = \sum_{j=1}^{A} P(a_j|z), \tag{43}$$

means the average number of purchased items of a user belonging to the latent class z. Investigating this value, the average purchase number of users belonging to the latent class can be compared.

On the other hand, from the equation

$$P(u_k|z) = \frac{1}{\sum_{k=1}^{U} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}, \boldsymbol{s}_{u_k})}\tilde{P}(z|u_k, \boldsymbol{x}_{u_k}, \boldsymbol{s}_{u_k}), \tag{44}$$

we have

$$\begin{aligned} P(u_k) &= \sum_{z \in \mathcal{Z}} P(u_k|z)P(z) \\ &= \frac{1}{U}\sum_{z \in \mathcal{Z}} \tilde{P}(z|u_k, \boldsymbol{x}_{u_k}, \boldsymbol{s}_{u_k}) \\ &= \frac{1}{U}. \end{aligned} \tag{45}$$

That is, the equal probability $1/U$ is allocated to every user in the user set in the proposed model.

V. An Application to Internet Portal Site for Job Hunting

In order to verify the effectiveness of our proposal, the proposed model is applied to the data of an internet portal site for job hunting.

A. Background Information

In Japan, many university students use several internet portal sites for job hunting. Their access to the portal sites and entry histories are saved in the database and we can use the data to learn the model. Usually, the third year university students start to think their job hunting before the summer. Some students participate in the internship of companies during a summer period. After the summer season, most of the third year students start job hunting activities in the fall. They create their account (ID and password) of the portal sites for job hunting and search job offers information from companies on the site.

It is desirable for the company managing such an internet portal site for job hunting to analyze the data and clarify the effective promotion strategy. Because the company managing this portal site gets a sales from the registering companies in proportion to the number of the entries by students, it is important task for them to increase the number of entries for various kinds of conpanies registering this site. The majority of entries are now concentrating on the popular companies so that it is desirable to increase the entries of students to companies with low ranking of entry number [12]. Although the decision tree and other predictive model can be applied to prediction of the number of entries [13],[14], it is difficult to model the relation between users and companies directly. In order to make an effective strategy, the probabilistic model to represent the entry behaviors of university students can be introduced.

On the case used in this study, the total number of students who have their account of a portal site is about 500,000 every year and that of companies is about 10,000. The company managing the portal site already introduced several web marketing tools, i.e. individual recommendation of companies to each student. It is desirable to derive a new marketing strategy from the different viewpoints.

B. The Training Data and Model Settings

By random sampling from students on the 2013 academic year, one year access and entry data are used for a training data. The training data is 5,000 students which were randomly sampled from the students living in Tokyo, thereby the number of entries in the training data is 301,401. The number of companies is $A = 9,065$. Letting the number of the latent classes be $Z = 10$, the tendency of students' preferences in each latent class is analyzed.

C. Model Fitting with EM Algorithm

Let the number of entries to the company a_j by all students be $N_a(a_j)$. After the training phase of the proposed latent class model, the estimated number of entries $\hat{N}_a(a_j)$ to the company a_j can be calculated by using the learned model. The mean square error between $N_a(a_j)$ and $\hat{N}_a(a_j)$ which is defined by

$$Err(N_a, \hat{N}_a) = \frac{\sum_{j=1}^{A} \left(N_a(a_j) - \hat{N}_a(a_j) \right)^2}{A}, \tag{46}$$

was 3.072×10^{-8} after model fitting with EM algorithm. Thus, the model fitting to the training data set is satisfactory.

TABLE I. The tendency of main attributes of students belonging to each latent class

latent class	main attributes of students	average number of entries	number of students
z_1	private university, apparel course	48.26	321.46
z_2	private and government universities, humanities course	51.15	554.00
z_3	private and government universities, science course	34.55	233.91
z_4	private university, humanities course	38.35	198.46
z_5	private university, science course	59.14	360.00
z_6	private university, humanities course	67.34	829.75
z_7	private university, humanities course	77.26	1294.90
z_8	private university, science course	63.03	446.01
z_9	private university, humanities course	45.96	360.71
z_{10}	private university, humanities course	49.00	400.80

Remark: Here, 'humanities' includes 'social sciences', 'and fine arts', and 'liberal arts'.

TABLE II. The tendency of companies applied by students belonging to each latent class

latent class	companies to which the students in each latent class apply
z_1	apparel and accessories
z_2	large enterprises
z_3	information processing and software industries
z_4	ceremony and tourism industries
z_5	game software and internet industries
z_6	advertisement and internet industries
z_7	general trading companies
z_8	large enterprises in semiconductor and electronic industries
z_9	large enterprises in food industry
z_{10}	real estate industry

D. The Students' Preferences of Entries to Companies

Table I and II show the result of analysis based on the proposed latent class model. In Table I, 'main attributes of students' means those of students who have high belonging probability $P(z|u_k)$. The 'average number of entries' is calculated by Eq.(43). The 'number of students' is defined by

$$N_u(z) = \sum_{u_k \in \mathcal{U}} P(z|u_k), \tag{47}$$

where $\sum_z N_u(z) = 5,000$.

Through Table I, the characteristics of students belonging to each latent class can be interpreted by the student attributes, i.e. 'private university or government university', and 'humanities course or science course'. From Table II, we can see that the students belonging to the same latent class tend to apply the similar companies with each other. The number of students belonging to the latent class z_7 is $1,294.9$ and the largest of all latent classes. The mean number of entries is also largest. We have clarified the fact that the general trading companies are very popular in Japan and many students who want to get a job in this industry are in job hunting actively. The main attributes of the students belonging to the latent class z_7 are private university and humanities course.

TABLE III. THE IMPACT WHEN 100 STUDENTS BELONGING TO THE LATENT CLASS INCREASE

latent class	A	B	C
z_1	4,925.9	4,724.3	201.7
z_2	5,114.4	4,971.3	143.1
z_3	3,454.4	3,420.5	33.9
z_4	3,835.3	3,406.8	428.5
z_5	5,913.8	5,760.8	153.0
z_6	6,734.4	6,516.8	217.6
z_7	7,725.5	7,402.5	323.0
z_8	6,303.0	6,004.2	298.8
z_9	4,595.9	4,301.6	294.3
z_{10}	4,900.1	4,771.2	129.0

Remarks:

A: the number of increased entries for all companies
B: the number of increased entries to companies with 5000-th or higher rank
C: the number of increased entries to companies with 5001-th or lower rank

E. Simulation with The Estimated Model

By using the proposed latent class model, we can calculate easily the impact when changing the number of students belonging to the latent class. This is the strong merit for the purpose of web marketing. We can specify the best target of latent class by the simulation analysis with the estimated model. Table III shows the impact when 100 students belonging to the latent class increase.

From this table, the more the students belonging to the latent class z_7 increase, the more the total number of entries increases. As mentioned in the previous subsection, the latent class z_7 is the segment of students who want to get a job in the general trading industry. However, it is not a good strategy to increase the size of latent class z_7 because the majority of entries are concentrating on the several popular companies. This strategy promotes popular concentration more.

It is the good strategy to increase the number of entries to companies with 5001-th or lower rank. This is because such companies are hoping the more entries from various kind of students. From the such viewpoint, the latent class z_4 is most important because 428.5 entries may increase when 100 students in z_4 increases. The impact of increase of students in z_4 is largest of all latent classes. Because we can make a list of departments of universities who have many students belonging to the latent class z_4, it is the good promotion to contact the departments with high priority. If new student members of this site from the departments in this list by holding briefing sessions, the increase of the entries to companies which are not popular can be expected.

As we demonstrate in this section, the proposed latent class model is effective for evaluating the impact by several changes of situation through the simulation. This simulation with the proposed latent class model can be used for strategic planning in practice.

VI. CONCLUSION

This paper proposed a new model whose probability parameter of latent variable means the rate of users with the same preference in market in order to use it for the simulation analysis of the market model. The effectiveness of our proposal has been verified by applying the proposed model to the data of an internet portal site for job hunting. The proposed latent class model is effective to clarify the impact by several changes of situation through the simulation.

The systematic procedure of market simulation by using the proposed model is one of the future works. In the demonstration of this study, the data set of 5,000 students living in Tokyo are used. The computational complexity of model estimation with the EM algorithm become huge when the number of users become large. In the case of the portal site for job hunting used as the case study in this paper, the number of students are about 500,000. The reduction of computational time and increase of the training data size are also future works. We are also trying to apply the proposed model to the analysis of purchase history data in an EC site.

ACKNOWLEDGMENT

The authors would like to thank Dr. Tota Suko and Dr. Shunsuke Horii, and all members of Goto laboratory for their support to our research.

This work was partly supported by JSPS KAKENHI Grant Numbers, 26282090 and 26560167.

REFERENCES

[1] T. Hoffmann, “Probabilistic Latent Semantic Indexing”, In Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval, pp.289-296, 1999.

[2] J. Magidson, J. K. Vermunt, “Latent class models for clustering: A comparison with K-means,” Canadian Journal of Marketing Research, Vol.20, 2pp.37-44, 2002.

[3] A. Bhatnagar, S. Ghose, “A latent class segmentation analysis of e-shoppers,” Journal of Business Research, Vol.57, pp.758-767, 2004

[4] F. Bassi, “Latent class factor models for market segmentation: an application to pharmaceuticals,” Statistical Methods and Applications, Vol.16, Iss.2, pp.279-287, 2007.

[5] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl: “GroupLens: An Open Architecture for Collaborative Filtering of Netnews,” *Proc. of The Conf. on Computer Supported Cooperative Work*, pp.175–186, 1994.

[6] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl.: “An Algorithmic Framework for Performing Collaborative Filtering,” *Proc. of 22nd Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.230–237, 1994.

[7] T. Hofmann, “Latent semantic models for collaborative filtering,” Journal ACM Transactions on Information Systems, Vol.22 Iss.1, pp.89-115, 2004

[8] K. E. Train, “Discrete Choice Methods with Simulation - Second edition,” Cambridge University Press, Cambridge, 2009

[9] A. P. Dempster, N. M. Laird, D. B. Rubin, “Maximum Likelihood from Incomplete Data Via The EM Algorithm”, Journal of Royal Statistics Society, Series B Vol.39, No.1, pp.1-38, 1977.

[10] G. McLachlan, T. Krishnan, The EM Algorithm and Extensions, Wiley-Interscience, 2007.

[11] D. Chen, D. Wang, G. Yu, F. Yu: “A PLSA-based approach for building user profile and implementing personalized recommendation”, in *Advances in Data and Web Management, Lecture Notes in Computer Science*, Vol.4505, pp 606-613, 2007.

[12] M. Hayakawa, K. Mikawa, T. Ishida, M. Goto, “A Statistical Prediction Model of Students’ Success on Job Hunting by Log Data,” The 14th Asia Pacific Industrial Engineering and Management Systems Conference (APIEMS 2013), 2013.

[13] J. R. Quinlan, “C.4.5: Problems for Machine Learning,” Morgan Kaufmann, 1993.

[14] C.M.Bishop, Pattern Recognition and Machine Learning, Springer, 2006.